Monitoring of Dyadic Conversations: a Social Signal Processing Approach

Yasir Tahir

School of Electrical & Electronic Engineering
Institute for Media Innovation

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirements for the degree of
Doctor of Philosophy

2016

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

....

Date Yasir Tahir

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Justin Dauwels for his consistent support and guidance since I joined Nanyang Technological University (NTU). I would also like to express my gratitude to my co-supervisor, Prof. Daniel Thalmann and Institute for Media Innovation for giving me a chance to work on this interesting research area.

I also owe my sincere appreciation to Debsubhra Chakraborty, Dr.Tomasz Maszczyk, and Smitha Velayil for their constructive suggestions and valuable contributions. Special thanks also to my previous colleagues Umer Rasheed and Sanath Sarda and the FYP students for helping me with different aspects of my work.

I would also like to express my gratefulness towards my family and friends, whose constant and never ending support have helped me complete this endeavor successfully!

Abstract

Monitoring of dyadic Conversations: a Social Signal Processing Approach

by

Yasir Tahir[1]

Supervisor: Assoc. Prof. Justin Dauwels[1]

[1] Nanyang Technological Univeristy, Singapore

This work presents a real-time system that analyzes non-verbal audio and visual cues to quantitatively assess sociometrics from on-going two-person conversations. The system non-invasively captures audio and video/depth data from lapel microphones and Microsoft Kinect devices respectively to extract non-verbal speech and visual cues. The system leverages these non-verbal cues to quantitatively assess speaking mannerisms of each participant. The speech and visual cues are incorporated as features in machine learning algorithms to quantify various aspects of social behavior including Interest, Dominance, Politeness, Friendliness, Frustration, Empathy, Respect, Confusion, Hostility and Agreement. The most relevant speech and visual cues are selected by forward feature selection. The system is trained and tested on two carefully annotated corpora, i.e., an Audio Corpus (AC) and Audio-Visual Corpus (AVC) comprising brief two-person dialogs (in English). Numerical tests through leave-one-person-out cross-validation indicate that the accuracy of the algorithms for inferring the sociometrics is in the range of 50% - 86% for AC and 62% - 92% for AVC. To test the robustness of the proposed approach, the audio data from both corpora are combined, and a classifier is trained on this mixed data set. Despite the significant differences in the recording conditions of the AC and

AVC, the accuracy for inferring sociometrics from this mixed data set is in the range of 51% - 81% which is reasonably high, therefore implying that the algorithms are robust to changes in the recording conditions. The proposed algorithms have low computational complexity. They can operate in continuous time, and yield sociometrics in real-time. Consequently, they can be implemented on real-life platforms. The term sociofeedback has been coined to describe systems of that kind, which are capable of analyzing conversations and providing feedback to the speakers based on their speaking patterns.

To obtain user feedback regarding practical implementation of sociofeedback system in realistic scenarios, the sociofeedback system was interfaced with a humanoid robot (Nao). This enabled the humanoid robot (Nao) to provide real-time sociofeedback to participants taking part in two-person dialogs. The sociofeedback system quantifies speech mannerism and social behavior of participants in an ongoing conversation, determines whether feedback is required, and delivers feedback through Nao. For example, Nao alerts the speaker(s) when the voice is too low or too loud, or when the conversation is not proceeding well due to disagreements or numerous interruptions. The user study about the Nao robot comprises two set of experiments. In the first sets of experiments, the participants rate their understanding of feedback messages delivered via the humanoid robot. They also assess two modalities to deliver the feedback: audio only and audio combined with gestures. In majority of the cases, there is an improvement of 10% or more when audio and gesture modalities are combined to deliver feedback messages. For the second set of experiments, the sociofeedback system was integrated with the Nao robot. The participants engage in two-person scenario based conversations while the Nao robot delivers feedback generated by the sociofeedback system. The sociofeedback system analyzes the conversations and provides feedback via Nao. Subsequently, the participants assess the received sociofeedback with respect to various aspects, including its content, appropriateness, and timing. Participants also evaluate their overall perception of Nao via the Godspeed questionnaire. Results indicate that the sociofeedback system is able to detect the social scenario with 93.8% accuracy, and that Nao can

be effectively used to provide sociofeedback in discussions. These results pave the way to natural human-robot interaction in a multi-party dialog system.

Another real world application of such a system that has been explored is non-verbal speech analysis to facilitate schizophrenia treatment. Negative symptoms in schizophrenia are associated with significant burden and functional impairment, especially speech production. In clinical practice today, there are no robust treatments for negative symptoms, and one obstacle surrounding its research is the lack of an objective measure. To this end, non-verbal speech cues are explored as objective measures. Non-verbal speech cues are extracted from schizophrenic patients and psychologist interviews. The interviews of the patients enrolled in an observational study on the effectiveness of Cognitive Remediation Therapy (CRT) are analyzed. The subjects comprise schizophrenic patients undergoing CRT treatment, and the control group consists of schizophrenic patients not undergoing CRT. Audio recordings of the patients are made during three sessions while being evaluated for negative symptoms over a 12-week follow-up period. In order to validate the non-verbal speech cues, their correlations with the Negative Symptom Assessment (NSA-16) are computed. The results suggest a strong correlation between certain measures in the two rating sets. Supervised prediction of the subjective ratings from the non-verbal speech features with leave-one-person-out cross-validation has shown a reasonable accuracy of 75-80%. Furthermore, the non-verbal cues can be used to reliably distinguish between the subjects and controls as supervised learning methods can classify the two groups with 69-80% accuracy.

Page intentionally left blank

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Non-verbal communication plays a vital role in understanding human behavior. The auditory and visual cues associated with the non-verbal communication are known as social signals. These social signals form a significant part of our everyday communication. Human behavior is very complex, and much attention has been given to the use of technology to facilitate in automated human behavior detection. This has given rise to research fields such as natural language processing and social signal processing. Natural language processing is the study of verbal aspects of speech, where in social signal processing the non-verbal aspects of communication are also investigated.

Our motivation for this work is to analyze dyadic conversations using a social signal processing based approach. To this end, we develop a system that can acquire audio-visual data, extract non-verbal cues and infer various social indicators namely *Interest*, *Dominance*, *Agreement*, *Politeness*, *Friendliness*, *Frustration*, *Empathy*, *Respect*, *Confusion* and *Hostility*. We believe that such a system has numerous applications to enhance the social intelligence of machines. Existing research also indicates the need for such a system. We have coined the term "Sociofeedback" for such a system.

In this thesis, we present the Sociofeedback system, feedback platforms and two real world applications for such a system.

We developed various applications which can be interfaced with the sociofeedback system and enable it to provide feedback on VoIP platform (Skype), Android platforms (Android phone/tablets), google glass and vuzix smart glass.

The first application is in the area of social robotics. Social robotics is the research field where human-robot interaction is studied in various social contexts. The ability to estimate the social state of the human with which it is interacting can help the robot to behave in a more socially aware manner. We interfaced the Sociofeedback system with a humanoid robot Nao. Utilizing the inference from the Sociofeedback system, Nao provided feedback to one or both speakers. The reason we chose social mediation scenario was to determine the following. The primary reason was to put the Sociofeedback system to the test. We trained the system on a corpus of dyadic conversations. The performance of this system in the social mediation context would reinforce our claims in our earlier work regarding the ability of the system to work in a real application. Secondly, we wanted to get an insight into the user opinion about a robot that was aware of their social state and could give them feedback about their social state. The participants rated the robot on a Godspeed questionnaire. They also rated the feedback provided by the robot.

Another real world application of a system similar to the Sociofeedback system that we have explored in our research is non-verbal analysis of schizophrenic patients' interviews. There is no system until date that can provide objective ratings for schizophrenia negative symptoms. Our objective is to determine the correlations between subjective negative symptoms ratings and non-verbal cues. We aim to exploit these correlation and implement a system that can provide objective ratings for negative symptoms criteria. We collected a corpus of schizophrenia patients and psychologist interviews over multiple sessions. The psychologists rated the patients on negative symptoms during the session. The non-verbal audio features were extracted from these interviews and we determined their correlation to the NSA

ratings. This is an ongoing project .Therefore we present the results obtained so far in this thesis. The final goal is to develop a system trained on the interview corpus can be used in an android or tele-medicine application to help with schizophrenic patient rehabilitation.

1.2 Sociometrics

Automatic analysis of human behavior from conversations has garnered a significant attention from psychologists, engineers, and data scientists because of the potential applications and numerous scientific challenges associated with it. Conversations consist of verbal and non-verbal cues, and experimental evidence from social psychology shows that social interactions are determined from display and interpretation of non-verbal cues [1]. This has led to a significant interest in devising computational techniques to analyze non-verbal cues from conversations. Non-verbal cues consist of tone, body gestures and postures, eye gaze, and facial expressions [2, 3]. Automatic analysis includes a three-step process [4]:

i. Data capture resulting in audio-video signals,

ii. Extraction of non-verbal cues,

iii. Inference of social interaction.

The domain that aims at this automatic analysis of human behavior from non-verbal cues is called Social Signal Processing (SSP) [4]. The potential applications of SSP span numerous areas such as communication, robotics, healthcare, education and business. We aim to automatically quantify social interactions from dyadic conversations and presents a system that can provide real-time sociofeedback through audio-video analysis of such conversations.

A comprehensive review of SSP [4,5] shows that research in this area varies in terms of the data acquisition methods, types of conversations, extraction methods, computational techniques, and the type of social interactions studied. Four categories of social constructs have been extensively studied viz. interaction management, internal states, personality traits, and relationships. Interaction management includes addressing and turn taking, with a focus mainly on identifying the addressee through speech, gaze, gesture, head pose, head movements, and facial expressions [4, 6–15]. Automatic detection of internal states such as interest and activity levels in multi-party dialogs has been explored from speech pitch [5, 16–18]. Similarly, detection of agreement has been investigated in meeting scenarios such as broadcast conversations [19–25]. In the majority of these studies, the proposed algorithms are evaluated on the annotations from the ICSI and AMI corpora [26, 27]. Dominance and related concepts such as emerging leadership have been investigated in a similar manner [28–30]. In addition, personality traits such as dominance and leadership have been investigated from speech energy, pitch rate, vocal control, and interruption. Leveraging on findings of social psychology that dominant people are more active and display significantly higher body movements [31–33], additional methods have been proposed to detect dominance in multi-party dialogs in an automated fashion by incorporating visual cues [34–36] such as body movement, gestures, expressions, and gaze. Table 1.1 provides a summary of the social indicators, corpora and the accuracies reported in the existing literature.

Several other studies examined modeling methods and automatic detection of social relations and social roles from speech recordings and short clips [4, 6–14]. It could also be noted that most of these studies inferred not more than one social signal, e.g., activity, dominance or agreement level [16–19, 34–36], which are primarily social attitudes. Relatively little work has been done towards the understanding of internal states such as conflicts, frustration, arguments, and boredom in face-to-face interactions (see [5]). The four aspects of social behavior can also be categorized into social attitudes (e.g., agreement, dominance), social feelings (e.g., envy, empathy, disgust), and social relationships (e.g., status, roles) [4]. We concentrate on social

Table 1.1: summary of the social indicators, corpora and the accuracies reported in the existing literature.

Bibliography	Social Indicator	Corpus	Results
[15]	Activity level	AMI	RR-34-55%
[16]	Emphasis detection	ICSI	Detection- 81-89%
[17]	Interest	AMI	Precision-0.75 Recall-0.55
[18]	Agreement	Broadcast conversation	F score- 63-92%
[18]	Disagreement	Broadcast conversation	F score-55-85%
[19]	Low Conflict	Political debates	F score-97.3%
[19]	High Conflict	Political debates	F score-87.1%
[20]	Agreement vs. Disagreement	ICSI	Accuracy-78%
[22]	Agreement/Disagreement recognition	Canal9 Database of Political Debates	Accuracy-64%
[24]	Agreement/Disagreement recognition	ICSI	Accuracy-87%
[28]	Influence Detection	AMI	Accuracy-71%
[29]	Dominance	AMI	Accuracy-84-91%
[33]	Most Dominant	AMI	Accuracy-82%
[33]	Least Dominant	AMI	Accuracy-86%
[35]	Leadership	broadcast conversations	F score-73-95%

attitudes and feelings (see Table 1.2). We will consider social relationships in future work.

The proposed real-time system can provide comprehensive feedback of social inter-

Table 1.2: Social indicators ("sociometrics") inferred in our study.

Social Attitudes	Social Feelings
Dominance	Frustration
Agreement	Empathy
Politeness	Confusion
Friendliness	Hostility
Respect	Interest

actions by inferring ten social indicators including *Dominance*, *Interest*, *Agreement*, *Politeness*, *Friendliness*, *Frustration*, *Empathy*, *Respect*, *Confusion*, and *Hostility* from non-verbal cues captured through audio and video. Our approach differs from the existing literature in the following ways:

Data collection using Kinect Sensor: Previous studies that explored the role of audio-visual features in detecting dominant people and emerging leaders in group conversations [34,37] utilized RGB cameras. We employed *Microsoft Kinect* sensors to extract visual features that provide reliable real-time face and skeleton tracking,

thereby enabling us to extract visual cues in real-time. To the best of our knowledge, 3D has only recently been used for visual cue extraction [38] to perform 3D mapping on RGB videos, and not on depth data. Although several corpora based on Kinect device such as CAM3D and BAVCD exist [39, 40], these are not related to human behavior studies.

Speech corpora: In order to test our proposed algorithms, we created two annotated speech corpora, one with audio data (audio corpus; AC) [41] and the second with audio-visual data (audio-visual corpus; AVC) [42]. In contrast to the existing studies [28] that manually extracted conversational parts from existing corpora such as AMI and ICSI [26, 43], we trained the system on the AC and AVC corpora with complete conversations. Our corpus is the first ever in this domain to consist of both RGB and depth visual data, and was obtained through audio-visual recordings of participants who were given scenarios for the conversations. Annotations were performed for various social constructs and exploited for supervised learning. We analyzed the feature set of each modality by means of these annotations, in order to assess the relevance of these feature sets for the prediction of social constructs.

Computational models: In [16, 18], dominance and interest in conversations are inferred by means of hidden Markov models (HMM); these studies demonstrate that the audio modality is predominant, and combining audio and visual information does not lead to better results. Moreover, rule-based, and rank-level analysis has been employed to determine the dominance and emerging leader. In contrast, our study employs supervised learning by leveraging on annotations from multiple judges, and we demonstrate that visual information can improve the detection of sociometrics. We allocated only one audio channel per speaker which allowed us to avoid speaker diarization and enabled us to determine features that involve interplay between speech patterns of both speakers.

Multiple social indicators: Unlike previous studies that inferred not more than one signal or one category of social construct i.e., social attitudes, we explored ten indicators namely *Dominance*, *Interest*, *Agreement*, *Politeness*, *Friendliness*, *Frustration*, *Empathy*, *Respect*, *Confusion* and *Hostility*; this portfolio of social indicators

provides comprehensive feedback on social behavior in terms of five social attitudes and five social feelings (see Table 1.2).

Real time sociofeedback system: Our system is able to provide feedback in real-time. The analysis duration is kept small in comparison to the data collection window, and that feedback is provided while the conversation is ongoing. For 1 minute of collected data, the analysis duration of our proposed system is around 7 seconds (see Fig. 2.14 and Table 2.13). In this manner, the system can efficiently compute feedback once the first minute of data has been collected. This is also because the pre-processing is fully automated; no manual selection of data segments is required in the proposed approach, in contrast to other studies [17, 44].

Correlation study: We conducted a correlation study to explore how the social indicators are related to each other. Our results indicate the existence of strong relations among the social signals, with positive indicators such as *Empathy* and *Politeness* positively correlated with each other; and negative indicators such as *Hostility* and *Frustration* positively correlated with each other, thereby indicating that positive emotions are negatively correlated with negative emotions.

In the majority of these existing studies [4, 6–14], the proposed algorithms are evaluated on the annotations from the ICSI corpus [26]. Typically, the corpus utilized for analysis and testing purposes is limited in scope and mostly contains a specific kind of dialog, e.g., broadcast shows [36]. In addition, the corpus is manually or automatically segmented based on activity levels as a pre-processing step, before the behavioral analysis can be conducted in most existing studies [17, 44]. Consequently, such methods cannot be used in real-time applications. It has been pointed out repeatedly in the literature that real-time systems for social signal processing would have many potential practical applications, yet few such systems have been developed so far [4, 5]. We aim to specifically address this shortcoming by concentrating on real-time implementations of social signal processing systems. Various research studies have proposed feedback systems to assist the dynamics of conversations by providing feedback to group members. This feedback is generated through the estimation of non-verbal cues (specifically, speaking time) associated with dominance

and measures of interaction; behavioral data is captured through wearable sensors and feedback is delivered on the meeting table or on cell phones [45–48]. For example, Meeting Mediator (MM) is a real-time portable system that provides feedback on mobile phones to enhance group collaboration by detecting social interactions captured through sociometric badges [46].

Other systems that determine speed dating or salary negotiation outcomes have

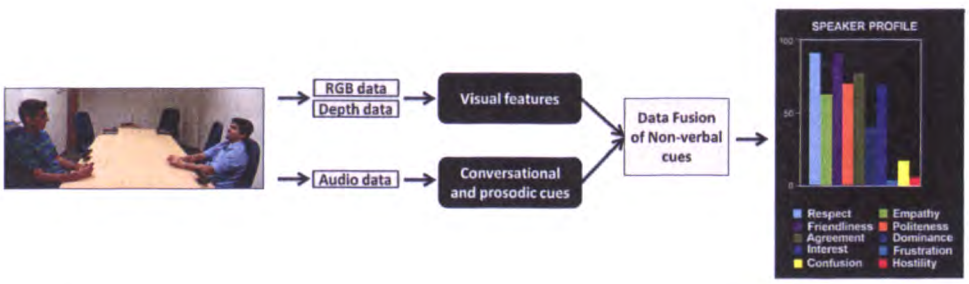


Figure 1.1: System Overview. The system records audio and video data, computes several speech and video cues, and from those features, computes levels of various sociometrics via multi-class classification. Each speaker is wearing a lapel microphone, and *Microsoft Kinect* sensors are placed in front of each speaker (see left figure).

been implemented [49]. However, these systems are unable to operate in real-time. It is noteworthy that the existing real-time feedback systems only provide low-level cues such as pitch and speaking percentage, and often infer only a limited number of social signals (often only one, at most three), limiting their use in practical scenarios where more diverse information might be required about the behavior of the speakers.

The motivation for this work arises from the above mentioned gaps; specifically, we aim to implement real-time systems that can provide comprehensive feedback on social behavior in dialogs on various platforms including VoIP, smartphones and smartglasses from non-verbal cues captured through audio and video.

1.3 Sociofeedback via Nao Robot

One of the key objectives of research and development in robotics is to design various robots that can assist humans in everyday domestic environments. Socially intuitive robots are utilized to communicate emotions and create social competencies [50, 51].There are many applications for such robots from shopping robots [52] and tour guides [53] to home assistance and care [54, 55] etc.

With increasing demand of robots for domestic environments, research on human-robot interaction (HRI) has gained more importance. In order to enhance human-robot interaction, the need for integration of social intelligence in such robots has become a necessity [56–58]. Socially intelligent robots should effectively engage with humans and maintain a natural interaction with them over extended periods of time.

Understanding of human behavior is essential for a robot to achieve social intelligence [59]. If a robot can understand the behavior of humans with whom it is interacting, then it can respond accordingly. HRI in multi-party dialogs [60] can be greatly improved if the robots are able to interpret the human behavior to some extent. Human behavior involves various patterns of actions and activities, attitudes, affective states, social signals, semantic descriptions and contextual properties [61]. A promising approach for human behavior understanding is to apply pattern recognition and automatically deduce various aspects of human behavior from different kinds of recordings and measurements, e.g., audio and video recordings [62].

In [41], we presented a novel approach towards comprehensive real-time analysis of speech mannerism and social behavior. We performed non-verbal speech analysis to analyze human behavior. Non-verbal speech metrics are a direct manifestation of human behavior, and play a vital role for the meetings to be pleasant, productive, and efficient [63]. By considering these low-level speech metrics, we quantified speech mannerism and sociometrics including interest, agreement, and dominance of the speakers. We collected a diverse speech corpus of two-person face-to-face conversations; it allowed us to train machine learning algorithms for reliable 5-level

classification of the sociometrics with speech metrics as input features. The classifier is able to detect social states of participants with accuracy of 84–86%. The combined metrics for speech mannerism and social behavior provided a clear picture of human behavior in dialogs. We investigate the scenario where the Nao robot communicates this information to the speakers.

In [64], we conducted a preliminary user study to investigate how sociofeedback could be provided via a humanoid robot (Nao). It is widely accepted that the combination of modalities and capabilities improves human-robot interaction. In our preliminary study, we investigated a variety of modalities. We provided users with sociofeedback in open-loop conditions. Specifically, the participants of the survey needed to assess basic feedback messages delivered by Nao, without actually participating in a conversation. The participants were then asked to assess sociofeedback messages delivered only via audio and also by a combination of audio and gestures. The user study confirmed the hypothesis that combining the two modalities of audio and gestures clearly helps the participants to identify the sociofeedback messages.

We extended our work from the open-loop to closed-loop scenario. The participants had a conversation, and the Nao robot provided feedback after wards. This feedback was derived from the speech mannerisms and sociometrics computed by the machine learning algorithms proposed in our earlier work [41]. In other words, we studied how the participants react to and evaluate this approach to providing feedback via Nao robot. We present the following contributions and novelties:

- We integrate a real-time sociofeedback system that analyzes nonverbal speech metrics to assess the social states of participants in a two-person conversation with a humanoid robot (Nao). The robot uses this information and provides appropriate feedback in real-time. Currently, we limit ourselves to four social states, namely normal, uninterested, overly talkative, and aggressive.

- We conducted two user studies in this research [64, 65]. The first user study with 20 participants (16 males, 4 females) compared audio, gesture and audio

with gesture modalities to provide feedback messages to speakers. In the second user study with 20 participants (17 males, 3 females). Each participant received sociofeedback via Nao for all the social states. Participants were then asked to evaluate several aspects of sociofeedback e.g., whether they *agree* with the feedback, or whether they *like* the feedback, whether they feel they received the feedback *timely*.

- We also investigated the overall experience of users about Nao. The participants were asked to rate the anthropomorphism, animacy, likability, perceived intelligence and perceived safety by means of a Godspeed questionnaire [66].

In summary, we made the following observations in our experiments. From our first experiment [64], we learned that the participants could clearly understand the feedback messages delivered by Nao using gestures along with audio. The ratings on Godspeed questionnaire were also significantly higher for the case when Nao used audio and gestures to deliver feedback message as compared to only using audio. In the second experiment we observed that the participants seemed to like receiving feedback from Nao robot and rated it very high on the Godspeed questionnaire. We also established that the sociofeedback system [41] can provide reliable feedback in real conversational scenarios with an overall accuracy of 93.8%.

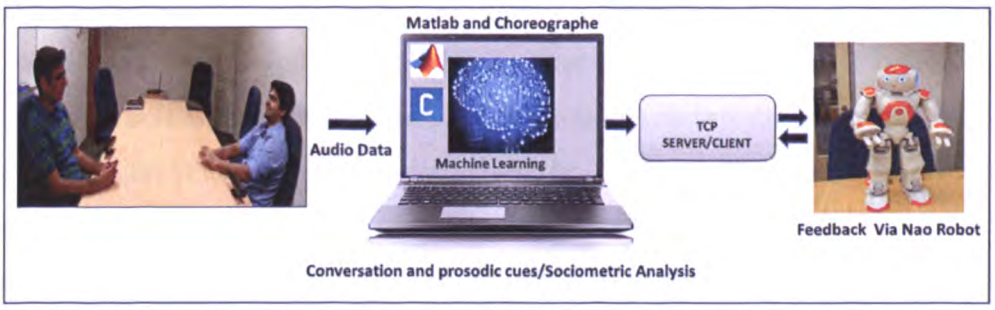


Figure 1.2: Feedback messages are determined from these three social indicators and from prosodic features. All these computations are performed in Matlab. The feedback messages are communicated from the computer to Nao via TCP/IP framework. Nao provides feedback by an audio message supported by gestures. The gestures are programmed in Choreographe.

In the recent past, many social robots have been designed for real world interactions, e.g., Kismet [67], Mel [68], Pearl [69], Robovie [70], Robota [71], and Paro [72]. Nowadays, social robots are successfully helping children in their social, emotional and communication deficits. Social robotics can help with the therapy of children with autism [73–75]. The roles of socially aware robots for autism therapy of children is reviewed in [76]. Similarly, social robots are being actively deployed in nursing homes for assistance of the elderly [72].

Apart from that, many application centric social robots are being deployed in domestic environments where the goal is to interact with humans as naturally as possible. The Human-Computer Interaction Institute (HCII) at Carnegie Mellon University (CMU) has developed an advisory robot that traces people's mental mode from a robot's physical attributes [77]. Similarly, Philips electronics created the iCat Research Platform which consists of a catlike face and a limbless body. iCAT has been deployed in [78] to investigate multi-party social interaction with a robot. CALO meeting assistant [79] assists participants in multiparty meetings. Furhat is a social robot developed to conduct research on multi-party dialogues with robots [80].

Many user studies have been conducted to assess how humans perceive robots in their specific roles. Such studies rate the human-robot interaction with respect to likability, perceived safety, anthropomorphism, animacy etc. For example, it was investigated in [81] how humans perceive affect from robot motion. It was shown in [82] that humans perceive different affects by observing different motions of the robot. The curvature and acceleration of robot motion were varied and their positive or negative affect on the participants were observed. Similarly, in [83] studies have been carried out to see if humans can identify emotions expressed by a humanoid using gestures. The results indicate that users can interpret expressions from different poses animated by a humanoid robot. In [84] Nao narrated a three minute story to a group of participants. The study investigated the effect of gazing and gestures on the persuasion of the robot, and provides evidence that gazing can significantly improve persuasion, however, incorporating gestures showed no significant difference in persuasion. In [85], experiments were carried out to understand whether a robot

can effectively modify its speech according to the speaker's behavior.

By contrast, our objective is to facilitate multi-party dialogs by introducing Nao as an observer, which can assess social state of participants, in real-time, and provide valuable feedback without having to provide any service or engage participants in any context-based conversation. To achieve this, we conducted a study to investigate, in detail, different aspects of human-robot interaction when Nao provides real-time sociofeedback to participants. To the best of our knowledge, no such study has been conducted yet.

1.4 Non-verbal Analysis of Schizophrenic patients' Interviews

Schizophrenia is a chronic and disabling mental disorder that often develops in adolescence and manifests itself through various symptoms. These symptoms can be classified broadly as positive, negative, and cognitive symptoms [86], [87]. Positive symptoms include one or more of the following: hallucinations, delusions, positive thought disorders, and repeated bizarre behavior [86]. Similarly, negative symptoms consist of at least two of the following: poverty of speech and its content, i.e., alogia, affective flattening, anhedonia-asociality, avolition-apathy, and attentional impairment [86]. In the literature, common cognitive deficiencies of schizophrenia such as failure to differentiate between relevant and irrelevant stimuli, lack of continued focus and attention, "an impaired capacity for abstraction, errors in syllogistic and analogical inference", and incoherent responses due to competing reactions [88], have been discussed. Although the pathogenesis of schizophrenia remains unclear till now, research so far has pointed to a strong genetic basis with estimates of heritability of risk at about 80% [89].

Current pharmacological treatments are effective in treating positive symptoms but have at most limited efficacy on negative and cognitive symptoms. Till now, their pharmacological treatments have been restricted to agents that act at the glycine

site of the *N*-methyl-D-aspartic acid (NMDA) glutamatergic receptor or a partial agonist, such as D-cycloserine [90–93]. However, the above clinical trials were mostly conducted in small, inpatient samples. When put to scrutiny using a larger sample of subjects with moderate to severe negative symptoms, Buchanan et. al [94] did not find the previous claims to hold up. In their "16-week double-blind, double-dummy, parallel group, randomized trial of adjunctive glycine, D-cycloserine, or placebo." viz. Cognitive and Negative Symptoms in Schizophrenia Trial, or CONSIST, they observed no significant differences in change in the Scale for Assessment of Negative Symptoms (SANS) scores and change in the average cognitive domain z scores between glycine and placebo subjects or D-cycloserine and placebo subjects. Murphy et al. reviewed the efficacy of the pharmacological agents on primary negative symptoms in [95]. They reported some drugs such as amisulpride to be only moderately effective in treating negative symptoms compared to the effect of conventional anti-psychotics on positive symptoms treatment. However, negative and cognitive symptoms in schizophrenia do contribute significantly to the disability seen in clinical practice [96]. Moreover, it has been seen in practice that negative symptoms and cognitive deficiencies of schizophrenia are closely related [97].

Identifying cognitive and sociological bio-markers will greatly aid the stratification and tailoring of treatment for patients with schizophrenia and can have the potential to be used as an objective way to gauge an individual's response to the treatment.

The advancements in speech processing research have paved the way for researchers

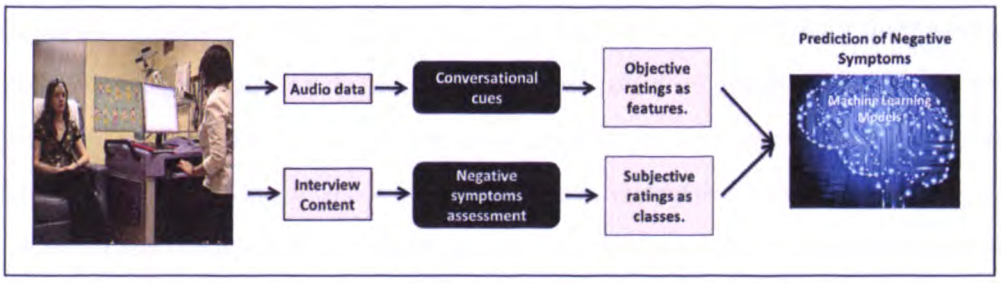


Figure 1.3: An overview of data acquisition and the analysis of subjective and objective features.

to investigate the assessment or diagnosis of mental disorders utilizing speech anal-

ysis. The most extensive work in this regard has been done for depression and suicide risk, and a review of existing work is provided in [98]. Prosodic features, formant features, source features and spectral analysis features can be extracted from speech and machine learning models can be trained on these features to predict the existence or the severity of mental disorders. In recent years, Interspeech competitions and data-sets have also motivated several studies towards a better understanding of mental disorders such as Parkinson and autism. Interspeech is a research competition held regularly since 2009 [99–105]. Each competition consists of multiple challenges, which encourage researchers to tackle diverse problems using speech features. The organizers provide the corpora and feature sets to be used for these challenges. In some of these studies speech or voice based cues are used to determine either the severity of the mental disorder or the relationship between speech cues and the mental disorder. Audio-visual emotion challenge (AVEC) is part of a workshop held at ACM-Multimedia conferences. It also includes depression recognition challenge [106, 107].

Speech impairments among schizophrenic patients are some of the key negative symptoms [108]. However, previous non-automated attempts to utilize the different aspects of speech and language as differentiators between schizophrenic and healthy individuals have had limited success. Although there existed some distinction in verbal fluency tasks between patients and healthy controls [109], other studies involving semantic boundary [110] or metaphor interpretation [111] reported no significant differences between the two groups. However, automated efforts based on speech deficiencies to distinguish patients and healthy controls have had greater success with the recent advancements in computer science and signal processing techniques. The ability of schizophrenic patients to express emotions was compared to that of a healthy group [112]. The results indicate that there exists a higher overlap between the measures of both groups for active primary emotions of anger and happiness as compared to sad emotion. Subtle differences in communication discourses were detected among patients, their first-degree relatives and healthy controls employing Latent Semantic Analysis (LSA) in [113]. LSA was again uti-

Table 1.3: Overview of studies concerning the use of speech analysis to assess the existence or the severity of Psychological disorders such as depression, suicide risk, schizophrenia, Parkinsons syndrome, and autism.

Psychological Disorder	References
Depression, Suicide Risk and PTSD	[98], [117], [118], [119], [106], [107]
Schizophrenia	[108], [109], [110], [111] [113], [114], [115], [112]
Parkinson's syndrome	[120], [121], [122]
Autism	[123], [124], [125]

lized to identify lack of semantic and phonological fluency, disconnected speech, and thought disorder in [114], and LSA and machine learning were used to analyze free-speech and predict the onset of psychosis respectively of high-risk youths in [115]. However, all the above methods are based on semantic analysis and natural language processing, with little attention to the restricted non-verbal cues display of patients suffering from negative symptom schizophrenia [116]. Non-verbal speech cues such as voice tone, volume, and interjections play a crucial role in human interaction and communication [2], and the display of such signals in patients can be used for both distinguishing them with healthy controls and developing specific and objective treatments. Table 1.3 presents the references relevant to research towards psychological disorders such as depression, schizophrenia, Parkinsons syndrome and autism. We provide a summary of few of the studies done in this regard.

The hypothesis that valuable information about depression can be determined from GMM of recorded speech was tested in [117]. Two corpora were used in this research. The first was the Mundt database originally collected for a depression severity study by Mundt et al. [126], involving both in-clinic and telephone-response speech recordings. This corpus consisted of thirty five subjects (20 women and 15 men, mean age 41.8 years). The interviews of these subjects were recorded at week 0,2,4, and 6, the subjects were assessed using HAMD scoring and QIDS assessment. HAMD [127] scoring is a clinical-rated test which provides a Htotal score, where QIDS [128] is a self-reported measure that provides a Qtotal score. Both use different weighting schemes to produce their total score. The best classification accuracy for Htotal and

Qtotal was 68.6% and 69% respectively.

The second corpus was the Black Dog database, which consists of audio-video data collected for a study conducted by the Black Dog Institute [126]. This database consisted of 60 participants with 30 subjects and 30 controls. The gender split was kept even, and there was one recording per participant. The best accuracy for Black Dog corpus was 63%. The results showed that the use of a speaker's average weighted variance which is a GMM based indicator enhanced the classification of both the presence and the severity of depression. Automatic classification of depression severity was also investigated in [119]. The researchers used a corpus of free-speech from subjects treated for depression over six weeks, along with standard clinical HAMD depression ratings. The results indicated that by mitigating nuisances and focusing on depression severity as a class, significant improvement in classification accuracy can be achieved.

Voice characteristics on a breathy to tense dimension were used as an indicator of psychological distress within semi-structured virtual human interview in [118]. The corpus, investigated in this work, was recorded in a wizard of Oz controlled scenario where a virtual human interacted verbally and non-verbally in a semi-structured manner with a participant. In total, 45 participants interacted with the virtual human. All participants who met requirements (i.e. age greater than 18, and adequate eyesight) were accepted. Their mean age was 41.2 years (27 male and 16 female). Significant differences were found between the voice quality of psychologically distressed participants and non-distressed participants. Classification accuracy of 75% was achieved for depression and 72.1% was achieved for PTSD.

As mentioned earlier, Interspeech competitions include mental disorders such as Parkinsons syndrome and autism to determine the role of speech analysis in understanding these disorders. Interspeech 2015 [105] had a Parkinsons condition sub-challenge. The Parkinson corpus included speech recordings of 50 people with PD and 50 healthy controls, 25 men and 25 women in each group. All the participants were Colombian Spanish native speakers. The age of the men with PD ranged from 33 to 77 years old (mean 62.2), the age of the women with PD ranged from

44 to 75 years old (mean 60.1). Acoustic, prosodic and glottal features on speech tasks such as syllable repetition, sentence reading and monologues were employed to determine the presence and severity of Parkinsons disease in [120]. The research questions addressed in this work were whether speech features could help in detecting Parkinsons disease and in estimating its severity. The first question had two possible outcomes (Parkinsons disease vs. healthy), and the second question had three outcomes based on unified Parkinsons disease rating scale (UPDRS). The researchers reported a recognition result of 82% when trying to differentiate between a normal speaker and a Parkinsons speaker, and 59% recognition result for UPDRS classification.

In another study [121] phone recognition based features were augmented with i-vector features to identify the degree to which a person suffers from Parkinsons disease. An improvement of 0.04 in the regression correlation coefficient over the baseline provided by Interspeech challenge [105] was reported. Similarly, an improvement over the baseline was reported in [122], where i-vector and functional segmental features, non-linear time series features, speech rhythm and automatic speech decoding based features were used.

Interspeech competition 2013 [103] had an autism sub-challenge. The Child Pathological Speech Database (CPSD) was used for this competition. It provided speech as recorded in two university departments of child and adolescent psychiatry, located in Paris, France. The data-set used in the Sub-Challenge contained 2.5 k instances of speech recordings from 99 children aged 6 to 18 years. Dyadic interactions of autistic children and psychologists were analyzed in [123], and speech characteristics such as monotonic speech, variable volume, atypical voice quality and slow speech rate were reported in autistic children. The research also reported that the psychologist's speech patterns changed based on his/her perception of the child's behavior. The correlations for child's features showed low correlations of 0.36 and 0.37, but the psychologist's features showed a higher correlation of 0.61 for both communication total and social interaction total.

An improvement in the detection of autism spectrum disorder (ASD) was reported

in [124], where researchers used fundamental frequencies and derived harmonic to noise ratio (HNR), shimmer and jitter from the reconstructed noise free speech signal. These features were used along with standard features such as energy, cepstral and spectral features. They employed regression to detect ASD and classification for classifying the sub-type. The results showed an improvement of 2.3% for ASD detection and 2.8% for sub-type categorization.

In [125] the relation between an autistic child's verbal response latency and his/her Electrodermal Activity (EDA) was studied. In some children, discriminative physiological patterns were found during short and long verbal response latencies. The classification results ranged from 50%-70%. This suggests that EDA signals contain information relevant to the amount of verbal response latency.

In summary, non-verbal speech analysis has been used to determine the existence and the severity of psychological disorders such as depression, Parkinson's syndrome and autism. The results are encouraging and avenues such as Interspeech and AVEC provide researchers with data and feature sets to further existing research.

In our work we have studied the correlations between speech features and negative symptoms from the negative symptoms assessment (NSA) scale, whereas in existing work speech analysis has mostly been used to determine the presence and/or the severity of mental disorders. In the case of schizophrenia, most of the existing research employs natural language processing. In our work, we have attempted to highlight the significance of non-verbal speech analysis for understanding negative symptoms of schizophrenia. We conducted this experiment in collaboration with the Institute of Mental Health, Singapore. Psychologists at IMH took the decision to use NSA-16 rating scale [129] for this experiment. We found the conversational features used in our work on sociometrics explained in Chapter 2 [41, 42] to be relevant to the speech related criteria in NSA-16 scale. Therefore, we have focused on the correlations between the NSA-16 criteria and conversational features extracted from the patient-psychologist interviews. Another recent study by Lavelle et al. [130] used non-verbal behavior of schizophrenic patients and their psychiatrists to understand the therapeutic relationship between them during meetings, but this study

observed the patient-psychiatrist interactions for a very brief period (6 minutes in total), and that too, only for one session. However, the consistency of the patients' behavior for different psychiatrists or over time may not be the same and the challenge to objectively understand and infer the behavior of patients suffering from negative symptoms schizophrenia utilizing their non-verbal speech cues still exists. We present the preliminary results on the relationship of such non-verbal speech cues extracted from audio recordings of patient interviews with the Negative Symptoms Assessment criteria [129]. The accuracy of machine learning algorithms with leave-one-person-out cross-validation technique in identifying the subjective ratings from objective cues is between 75-80%. We also report how the non-verbal cues were applied as features in machine learning algorithms to differentiate between subjects and controls, the former group suffering from a greater degree of cognitive impairments of schizophrenia than the latter. The accuracy of machine learning algorithms with leave-one-person-out cross-validation technique in attributing the features to subject cases and controls is between 69-80%.

Chapter 2

Sociometrics

In this chapter we present the details of our work on sociometrics. In section 2.1, we describe the hardware used to acquire data. In section 2.2 we elaborate on the audio and audio-visual corpora collection and annotation. In section 2.3, we explain the non-verbal audio-visual features extracted from dyadic conversations. In section 2.4, we present the correlation analysis of audio-visual features with social indicator annotation, and the correlations between the social indicator annotations. In sections 2.5, 2.6,and 2.7, we present the leave-one-person-out cross-validation results of various machine learning algorithms. Lastly in section 2.8, we detail the computational complexity of our proposed system.

2.1 Hardware

2.1.1 Audio

The hardware consists of easy-to-use portable equipment for recording the conversations. The audio hardware simply involves lapel microphones for each speaker and an audio interface device such as H4N recorder to allow multiple microphones to be interfaced with the computer and perform recording simultaneously (Fig. 2.1). We use lapel microphones to record conversations, these microphones are easy to

wear. Speech quality is also greater as the microphone is close to the mouth of the speaker.

Figure 2.1: H4n recorder on the left and lapel microphone on the right.

2.1.2 Video and Depth data

Kinect cameras are used to capture video and depth data of each speaker. Kinect cameras allow facial and body recognition of individual speakers in real-time (Fig. 2.2).

Figure 2.2: Kinect sensor.

Figure 2.3 illustrates the physical capabilities of Kinect sensor. On the left it shows the vertical and horizontal angles which Kinect sensors can capture, while on the right it shows the suitable distance from Kinect. This distance is necessary to detect skeleton properly.

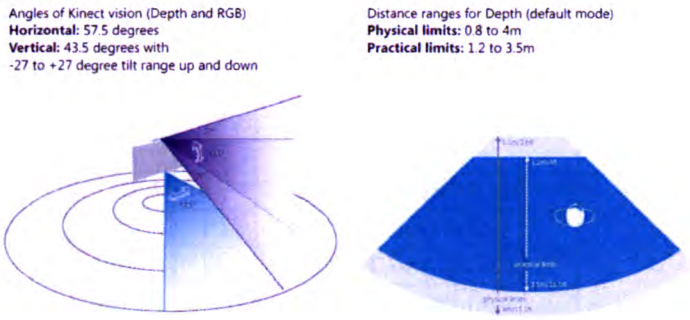


Figure 2.3: Limitations of Kinect sensors are shown in this figure.

2.2 Collected Corpora

We collected two different speech corpora, i.e., the Audio Corpus (AC) and Audio-Visual Corpus (AVC). Participants of both corpora are students of Nanyang Technological University (NTU). The Nanyang Technological University Institutional Review Board approved this study and experimental paradigm. All participants gave informed written consent and received monetary remuneration for their participation. The AC contains 150 two-person conversations, each at least 2.5–3 minutes long, recorded from a total of 22 participants (17 males and 5 females). The age of the students ranged from 18 to 30 years. On the other hand, the AVC contains 98 two-person conversations, each at least 1 minute long, recorded from a total of 21 participants (16 males and 5 females). The age of the students ranged from 18 to 30 years. The topics of conversations ranged from discussions of assignments, projects of students, to social and political views. They were also presented with problematic situations such as conflicts and disagreements, periods of boredom, aggressive behavior, or poorly delivered speech (e.g., low voice or fast pace). The length of each recording of our corpora is relatively long (at least one minute) as compared to existing corpora [12, 14]. The number of recordings for the different conversation topics in each corpus is listed in Table 2.1. In Appendix .1 we provide scripts of a few sample conversations to provide an idea of the conversations in the corpora. These scripted conversations were used for experiments to test the sociofeedback system, detailed in 4.5.

Table 2.1: Number of recordings for the major conversation topics in each corpus.

Topic	AC	AVC
Research	25	5
Career and Studies	23	23
Entertainment	14	13
Shopping and Lifestyle	22	19
Food and Culture	9	6
Social Issues	27	17
Sports	11	5
Travel	5	8
Religion and Politics	14	2
Total	**150**	**98**

2.2.1 Data Collection Procedure

The basic arrangement for data acquisition was kept similar for both corpora. Some observations from AC did affect our acquisition procedure for AVC. We collected the AC corpus in an uncontrolled manner, i.e., both people in the conversations were participants. The AVC corpus was recorded in a more controlled manner: one person in the conversations was a control, whose task was to facilitate the participant in conducting the dialogs for different social scenarios. The duration was 2.5–3 minutes in AC, but was reduced to 1 minute for AVC, as participants in the AC reported that the duration of the conversations was too long, and it was challenging for them to keep acting according to their roles in the conversation. This in turn enables us to investigate whether it is possible to provide feedback within shorter duration.

The procedure for data collection is listed below:

1. First, we make the recording system ready for data collection. We saved the speech in a 2-channel audio .wav file (one channel for each speaker), so that we can easily detect who is speaking at any given time. In the AVC, we recorded video/depth data for each participant by means of *Microsoft Kinect* devices in addition to audio.

2. The participants sat about 1.5 m apart, so that there was no interference from the other participant in the audio recordings. We placed the *Microsoft Kinect* devices in front of each participant (about 1.2 m apart), so that we could accurately extract the participant's skeleton and face from the depth and RGB data.

3. We briefed the participants about the experiment, and asked them to act naturally. The two participants were asked to agree on a topic for each subsequent discussion. As mentioned before, the topics of discussion ranged from casual small talk to heated debates on sports or politics (see Table 2.1). The topics were selected carefully in order to evoke a variety of behaviors.

2.2.2 Annotation Protocol

Each recording in the corpus was annotated by multiple people ("judges"), each assessing a subset of the corpus. There were 17 annotators for the AC corpus and each recording was annotated by a minimum of 4 and a maximum of 5 annotators. For the AVC there were 10 annotators and each recording was annotated by 5 annotators. For each recording in the AC, the judges completed a questionnaire related to speaking mannerisms (see Fig. 2.4) and behavioral aspects of each participant (see Fig. 2.5). In the latter figure, we show as an illustration the questions related to three social indicators, including *Interest*, *Dominance*, and *Agreement*; the real questionnaire had similar questions for each social indicator.

For example, if a participant seemed bored to the annotator, the latter would assess the interest level as "low"; in contrast, if the participant seemed excited, the annotator would quantify the interest level as "high". The responses ranged from 1 (low) to 3 (high). To simplify the annotation process, for the AVC the judges only annotated the social indicators, by answering questions similar to the ones shown in Fig. 2.5.

To assess the variability among the different annotators, we computed the standard deviation of the annotations. In Table 2.2 we list the standard deviation values for the 10 different sociometrics. As can be seen from Table 2.2, the standard deviation values are relatively low, and therefore, the annotations are reasonably consistent among the different annotators. The annotations for the AVC are on average more consistent than for the AC; this is probably due to the fact that also video footage is provided to the annotators in the AVC, in addition to audio, which may facilitate the annotation process.

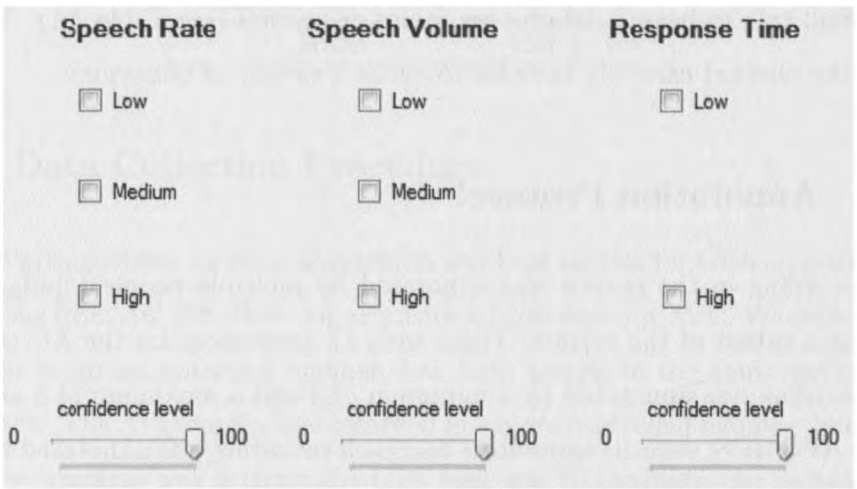


Figure 2.4: Questionnaire for speech mannerism assessment.

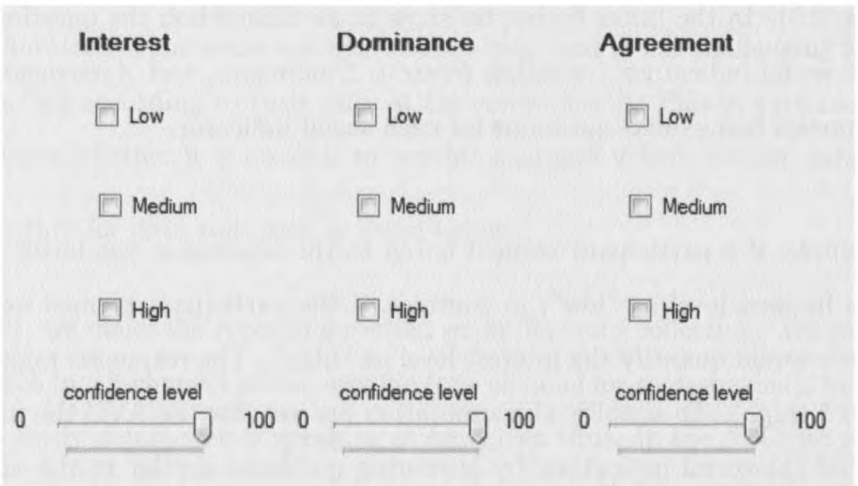


Figure 2.5: Questionnaire for sociometric assessment.

Table 2.2: The standard deviations among the annotations for each social indicator. The middle column shows the values for the audio corpus (AC), and the last column shows the values for the audio-visual corpus (AVC).

Social indicator	AC	AVC
Agreement	0.44	0.48
Dominance	0.50	0.24
Interest	0.54	0.39
Politeness	0.36	0.51
Friendliness	0.50	0.50
Frustration	0.43	0.33
Empathy	0.55	0.47
Respect	0.54	0.48
Confusion	0.36	0.34
Hostility	0.24	0.39
Average	**0.45**	**0.41**

2.3 Feature Extraction

Here we briefly review the speech and visual cues considered in this study (see Table 5.1).

2.3.1 Non-Verbal Speech Cues

Non-verbal speech cues can be divided into conversational and prosodic cues. In the following section, we briefly describe the conversational and prosodic cues.

2.3.1.1 Conversational Cues

In order to compute the conversational features, we first performed speech detection by means of a Hidden Markov Model (HMM) that employs energy-independent features [131]. Once the audio signals were segmented into periods with speech and without speech, we computed the following conversational cues: the number of natural turns, speaking percentage, mutual silence percentage, turn duration,

Table 2.3: List of conversational, prosodic and visual features.

Category	Features
Conversational	
Speaking duration	Speaking % , mutual silence, Difference in speaking %, overlap, response time
Speaking turns	Natural turns, turn duration
Interruption	Interruptions, failed interruptions
Interjection	Interjection, speaking interjection
Prosodic	
Frequencies	Larynx frequency (F0), formant (F1, F2, F3)
MFCC	Mel-frequency cepstral coefficients
Amplitude	Mean volume, max volume, min volume, entropy
Visual	
Postures	Upright, hunched forward, Leaned back, posture changes
Head movement	Nodding, Sum of vertical/horizontal head movements
Gestures	Gesture count, % of gestures as compared to other speaker.
Head pose	Straight, downward, sidewards

interjections, interruptions, failed interruptions, and response time, as illustrated in Fig. 2.6.

2.3.1.2 Prosodic Cues

We considered the following prosodic cues: amplitude, larynx frequency (F0), formants (F1, F2, F3), and mel-frequency cepstral coefficients (MFCCs); these cues were extracted from 30 ms segments at a fixed interval of 10 ms [41]. These cues fluctuate rapidly in time. Therefore, we computed various statistics of these cues over a time period of several seconds, including minimum, maximum, mean, and entropy.

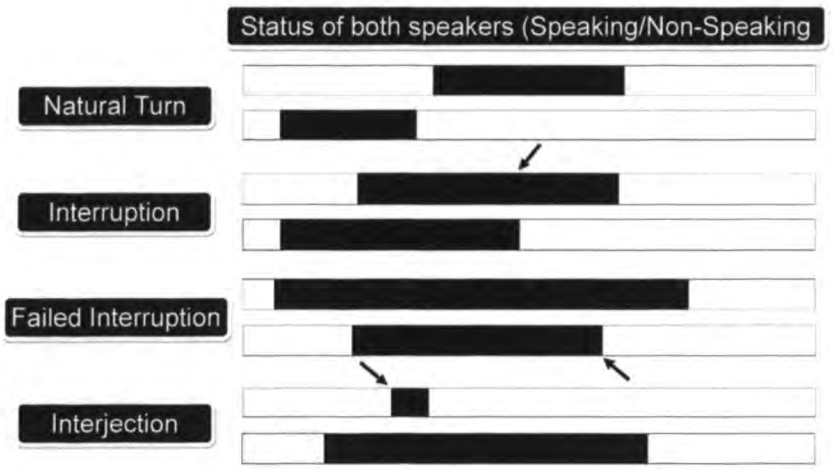


Figure 2.6: Illustration of natural turn, interruption, failed interruption and interjection. Periods of speaking and non-speaking are indicated in black and white respectively.

2.3.2 Visual Cues

We incorporated postures, nodding, hand gesture, and head pose in our analysis. It was challenging to extract the gaze of participants, as they were seated at approximately 1m distance from the *Microsoft Kinect* device. They also moved freely to express certain social states; therefore, we computed the head pose as a substitute for the gaze. We acquired RGB video and depth information from the speakers by means of two *Microsoft Kinect* devices (one for each speaker). In order to extract visual cues, the speaker's face and body were automatically detected first, as illustrated in Fig. 2.7.

Posture: Postures of each participant were classified into three basic sitting postures: hunch forward, upright, and lean back. We evaluated the percentage of time each participant remained in a particular posture, along with the total number of posture changes. Postures contribute strongly to understanding the body language and mood of the speaker. For example, an upright and erect posture indicates con-

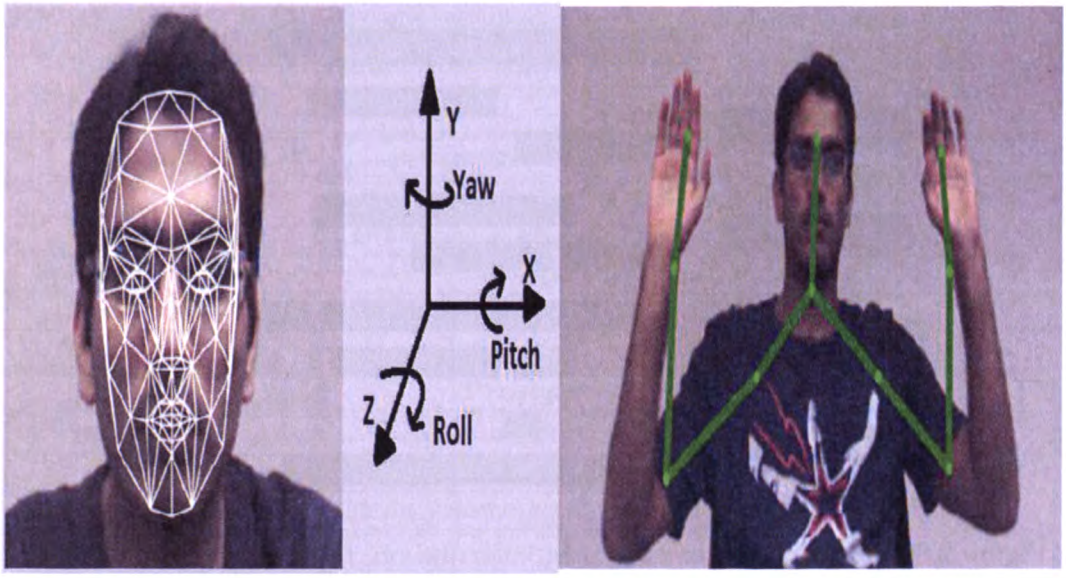


Figure 2.7: Face and skeleton detection from RGB-depth data acquired via *Microsoft Kinect* device.

fidence and positive thoughts in the speaker, whereas a slouching posture indicates doubtful and negative thoughts [132]. Since the skeleton acquired from *Microsoft Kinect* SDK [133] allowed us to track several points on the face and body of the speaker, we selected relevant points on the head (X_{head}, Y_{head}, Z_{head}) and neck (X_{neck}, Y_{neck}, Z_{neck}). By means of the angle θ between these points, we quantify the posture:

$$\theta = \tan^{-1} \frac{Y_{\text{head}} - Y_{\text{neck}}}{Z_{\text{head}} - Z_{\text{neck}}}. \tag{2.1}$$

The algorithm monitors the angle θ. Specifically, a posture change is registered whenever the angle changes and then remains in a particular range for at least 20s.

Head Pose: We utilize the face detection module of *Microsoft Kinect* SDK; it provides the head position in right-hand coordinates, in addition to yaw, pitch, and roll angles (see Fig. 2.7). We identified the head pose from the yaw and pitch angles of the detected face. For the sake of simplicity, the head pose of each participant was classified as straight, sidewards, or downward. The pitch value is negative when the person is looking down, and positive when the person is looking up. If this value remains positive or negative for at least 10s, the algorithm registers it as a

head pose. We use the sign changes in yaw values to determine the sidewards head pose; whenever the sign changes from positive or negative or vice versa, the person has moved his/her head sidewards. We also quantified the duration for which the participants remained in a particular head pose.

Nodding: Nodding is intuitively a good indicator of agreement and disagreement between the speakers. Nodding is commonly of two types: Yes and No. We registered consecutive head movements as nodding in the algorithm. Vertical head movement was determined by the pitch value (see Fig. 2.7): the pitch value is negative when the person is looking down and positive when the person is looking up. Vertical nodding ("Yes") is typically associated with repeated changes in sign of the pitch. When such changes occur during a period of at least 4s, the algorithm registers the occurrence of vertical nodding. Similarly, horizontal nodding ("No") corresponds to repeated changes in the sign of the yaw (see Fig. 2.7).

Hand Gesture Detection: We did not identify specific gestures, rather we quantified how often the speaker made hand gestures. Specifically, we defined the number of hand gestures and the relative percentage of hand gesture made by both speakers as visual features. We calculate the distance between the right hand coordinates $(X_{\text{rh}}, Y_{\text{rh}}, Z_{\text{rh}})$ and the left hand coordinates $(X_{\text{lh}}, Y_{\text{lh}}, Z_{\text{lh}})$:

$$\text{Distance} = \sqrt{(X_{\text{lh}} - X_{\text{rh}})^2 + (Y_{\text{lh}} - Y_{\text{rh}})^2 + (Z_{\text{lh}} - Z_{\text{rh}})^2}. \tag{2.2}$$

The algorithm monitors the change in distance between the hands, if this change is greater than 0.1m the algorithm recognizes it as a gesture. This threshold value has been experimentally determined.

2.4 Correlation Analysis

This section presents the correlation analysis of sociometric indicators for each corpus. First, we explore the interrelation between the social indicators themselves in the two corpora AC and AVC; then we provide examples of the relation between the social indicators and the nonverbal audio-visual features for the two corpora. The purpose of conducting this correlation analysis is to understand the relation between the indicators, in order to gain more insight into the hidden latencies among them and to express these relations in an engaging, visual manner. This analysis can help us identify the independent indicators and focus our attention on them. In addition, the study of the relation between indicators and audio-visual features can aid us to identify the most relevant non-verbal cues for each social indicator. Finally, these studies can uncover new relations between the indicators or features, which may be of interest to sociologists.

2.4.1 Relation between sociometric indicators

As mentioned earlier, the corpora AC and AVC have been annotated for ten social indicators viz. *Politeness*, *Friendliness*, *Frustration*, *Empathy*, *Respect*, *Confusion*, *Hostility*, *Agreement*, *Dominance* and *Interest*. The relationship between these indicators is shown in Fig. 2.8 and Fig. 2.9. This type of a figure is called a "Schemmaball" [134]. This relation is based on the linear correlation coefficient between the indicators, which has been calculated for all possible pairs of indicators. The linear correlation is a measure of the degree of association in a linear sense between two variables. We calculate the linear correlation value $\rho_{X,Y}$ between the annotation x_i of the ith recording of one social indicator, say *Politeness*, with the annotation y_i

Table 2.4: Group of curves based on correlation coefficient values.

Correlation coefficient values	Group name
[0.00, 0.33]	weak
(0.33, 0.66]	moderate
(0.66, 1.00]	strong

of another social indicator, say *Friendliness*, as follows:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.3}$$

where $\bar{x}$ and $\bar{y}$ denote the mean annotation value for all recordings of respective indicators. We have categorized these correlation coefficients into three groups, as explained in Table 2.4.

These categories have been represented in Fig. 2.8 and Fig. 2.9 as the thickness of the curves joining any two indicators, where the thickest curve represents the "strong" category and so on. Moreover red curves indicate negative correlation (values $\in [-1.00, 0.00)$), whereas green curves indicate positive correlation (values $\in [0.00, 1.00]$).

As can be seen from Fig. 2.8 and Fig. 2.9, there are strong positive correlations between *Hostility* and *Frustration* in both the corpora, whereas there are strong to moderate positive correlation between *Politeness* and *Respect*, between *Politeness* and *Friendliness*, and between *Empathy* and *Friendliness*. Similarly, there is strong to moderate negative correlation between *Politeness* and *Hostility*, between *Hostility* and *Respect*, and between *Politeness* and *Frustration*. All these relationships reaffirm the common perception that positive emotions such as *Politeness*, *Friendliness*, *Empathy* and *Respect* are positively inter-related. Similarly, negative emotions such as *Frustration* and *Hostility* are positively related to each other and on the whole are negatively related to the positive emotions in both corpora. These relations suggest redundancies among the indicators, and implies that we can reduce the number of

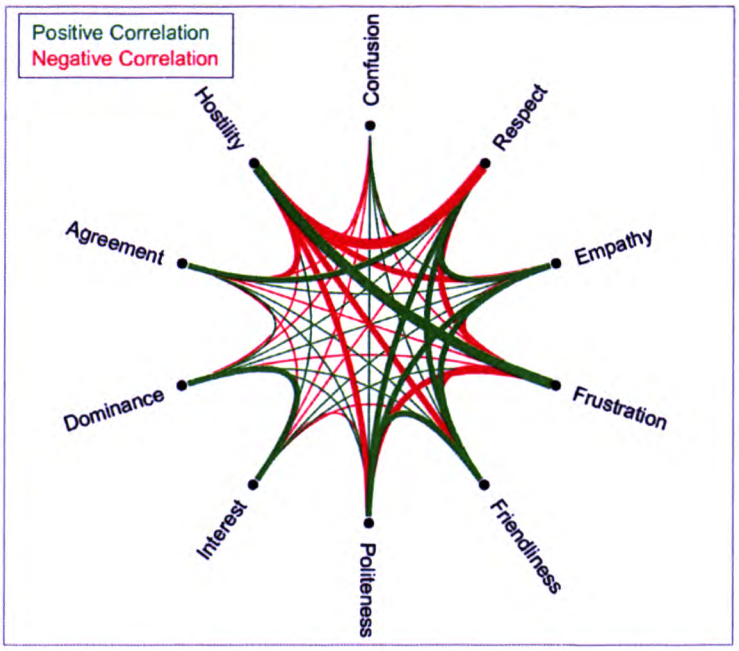


Figure 2.8: Correlation between indicators for AC corpus.

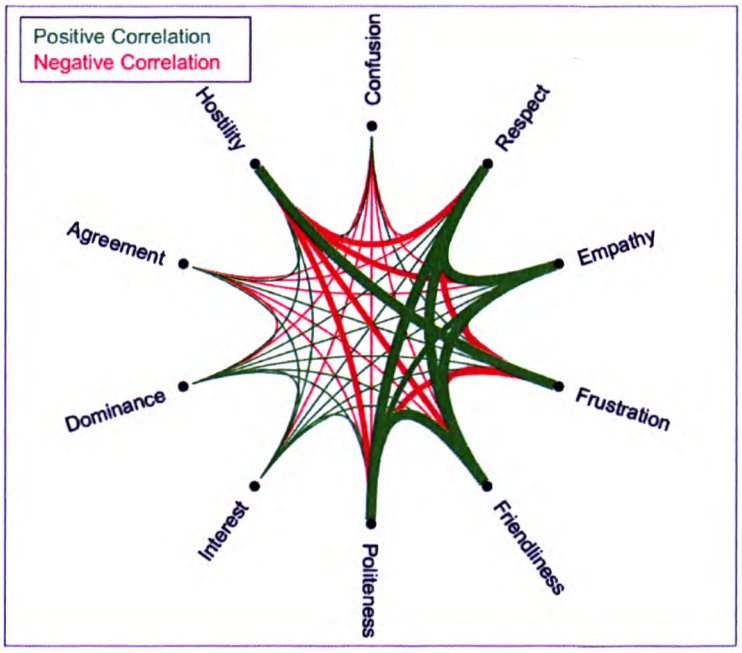


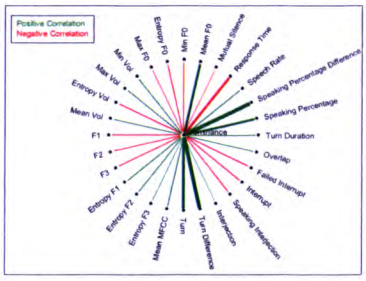
Figure 2.9: Correlation between indicators for AVC corpus.

indicators for the proposed real-time sociofeedback system, and yet provide feedback for a multitude of high-level social signals. From Fig. 2.8 and Fig. 2.9 it can be noted that the indicators *Confusion*, *Agreement*, *Dominance* and *Interest* are weakly correlated with each other and other indicators, and thus we can consider them as independent indicators. Also, another interesting observation is that *Confusion* did not appear to be correlated with any other indicator. It is also noteworthy that the correlations for the AC and AVC are very similar, suggesting that the annotations are stable across the two corpora. Moreover, the positive correlations are stronger for the AVC. This may be explained by the fact that the annotations are more consistent for the AVC (see Table 2.2), leading to more significant correlations between the indicators.

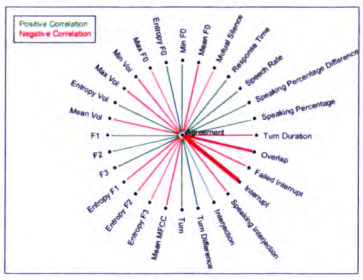
2.4.1.1 Relation between sociometric indicators and non-verbal audio-visual features

From the AC corpus, we extract various non-verbal audio features for each speaker (see Table 5.1). We extract the same audio features for the AVC corpus, in addition to non-verbal visual features (see Table 5.1). In Fig. 2.10 we depict the correlations between the non-verbal cues and the social indicators. This type of figure can be compared to a wheel; the social indicator is placed at the center as the hub of the wheel, whereas the non-verbal cues are placed on the wheel rim; the spokes of variable thickness represent the correlations between the social indicator and non-verbal cues. As mentioned earlier, green and red spokes depict positive and negative correlations respectively. In this section we provide two figures for each corpus, the other figures are included in Appendix .2.

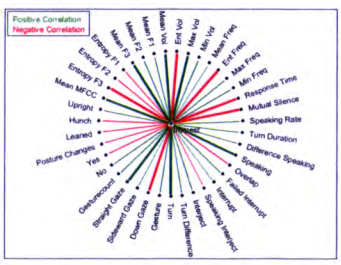
In Fig. 2.10(a), the inter-dependence between the social indicator *Dominance* and audio features from the NVAC dataset are shown. The social indicator *Dominance* has a strong positive correlation with the nonverbal features *Turn Difference* and *Speaking Percentage Difference*, which is understandable as a "dominant" person



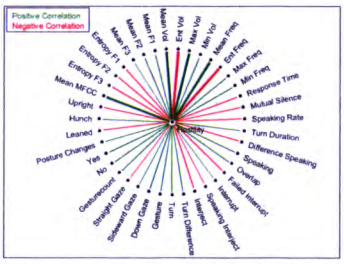
(a) Correlation between *Dominance* and audio features from AC corpus.



(b) Correlation between *Agreement* and audio features from AC corpus.



(c) Correlation between *Interest* and audio-video features from AVC corpus.



(d) Correlation between *Hostility* and audio-video features from AVC corpus.

Figure 2.10: Correlation between social indicators (*Dominance*, *Interest*, *Agreement*, and *Hostility*) and audio-video features for AC and AVC corpora.

will speak more and hence will take more speaking turns in conversations. Also, it has a moderate negative correlation with *Response Time*, which we can interpret as a "dominant" person will respond quickly and will have less *Response Time*. Fig. 2.10(b) depicts the relationship between the social indicator *Agreement* and the audio features from the audio corpus (AC). The social indicator*Agreement* has a strong negative correlation with *Interrupt*, indicating a conversation in which both speakers are in "agreement" with each other will have less interruptions. In Fig. 2.10(c), the relation between the social indicator *Interest* and the audio and video features from the audio-visual corpus (AVC) is shown. For the audio-visual corpus, the social indicator *Interest* is not strongly correlated with any of the features, but is at most moderately correlated with a few of them. For example, it has moderate positive correlation with the feature *Straight Gaze* and has moderate negative correlation with *Down Gaze*. It shows that a person with greater "interest" in the conversation will mostly look straight into the other person and less downward. Similarly, *Interest* has moderate negative correlation with *Mutual Silence*, indicating an "interesting" conversation will have less silence between the two participants. Fig. 2.10(d) shows the relation between the social indicator *Hostility* and the audio-visual features from AVC. The social indicator *Hostility* has moderate positive correlation with *Interrupt* and *Overlap* which indicates that in a conversation where the speakers are "hostile" towards each other will have more interruptions and speaking overlaps.

2.5 Audio Corpus (AC)

In this section we describe the proposed automated system to infer speech mannerisms (Section 5.1) and social indicators (Section 5.2) from non-verbal audio cues. The algorithms are trained and tested on the Audio Corpus (AC).

Table 2.5: Detection accuracies for the speech mannerisms of high/low volume and high/low speech rate. Confusion matrices are provided in the third column.

Speech Mannerism	Audio Feature	Detection (%)	Confusion Matrix		
			L	M	H
Speaking loudly/quietly	Volume	90	92%	8%	0%
			7%	88%	5%
			0%	10%	90%
			L	M	H
Speaking too fast/too slow	Speech Rate	84	95%	5%	0%
			19%	72%	9%
			0%	6%	94%

Table 2.6: Detection accuracies for the speech mannerisms of long response time. Confusion matrix is provided in the right column.

Speech Mannerism	Audio Feature	Detection (%)	Confusion Matrix	
			L	H
Slow Response	Response Time	96	90%	10%
			2%	98%

2.5.1 Speech Mannerisms

From the speech cues, we evaluate a variety of speech mannerisms: the speech volume is excessively low/high, the speech rate is excessively low/high, and the speaker is taking an excessive time to react. The system screens important speech cues for every speech mannerism and checks whether any cue is anomalous. From the scores given by the judges, we can characterize speech mannerisms. For example, when the score for "This person was too loud" is 2.5 or higher, the speaker is considered to be too loud.

In order to recognize such mannerisms from the speech cues, we selected certain thresholds. For instance, a volume level below 30 dB is classified as too low, and a volume level of over 80 dB is classified as too high; otherwise, the volume level is considered as normal. In this manner, we implemented a threshold-based scheme for automated classification of speech mannerisms. In Table 2.5 and 2.6 we have

listed the detection accuracies for each speech mannerism, along with the confusion matrices. Overall, the results suggest that basic speech mannerisms can be detected accurately from speech cues.

2.5.2 Inferring Social Indicators from Non-verbal Audio Cues

Here we describe the proposed system that infers social indicators from non-verbal audio cues (see Fig. 1.1). First we determine the most relevant non-verbal speech cues for these social indicators by feature selection, next we infer the social indicators from the selected speech cues by multi-class classification.

2.5.2.1 Feature Selection

Since it is computationally expensive to calculate numerous audio features, especially in the context of real-time applications on smartphones or smartglasses, we applied the correlation coefficient technique [135] to determine the most salient non-verbal speech cues for inferring the sociometrics. We applied the forward selection [136] approach to select the most relevant features for classification. In forward selection, features are progressively incorporated into increasingly large subsets, where we start with the most salient feature and then increase the subset until we obtain the best feature set. To determine the most significant feature set we apply leave-one-person-out cross-validation on the training data. The features are ranked based on their correlation coefficients and then the cross-validation is repeated with increasing feature subset until we get the best accuracy. In Fig. 2.11 we show the plot of social indicators against the audio features, where the values show the frequency of occurrence of the features in the selected feature set, thus showing their importance in obtaining best accuracy for the corresponding social indicator.

In Section 2.4 and Appendix .2, we presented "wheel" diagrams showing correlation of features for each social indicator, whereas Fig. 2.11 shows the frequency of occurrence of features in determining the best accuracy for each social indicator. We can see some similarities between the wheel diagrams presented earlier in

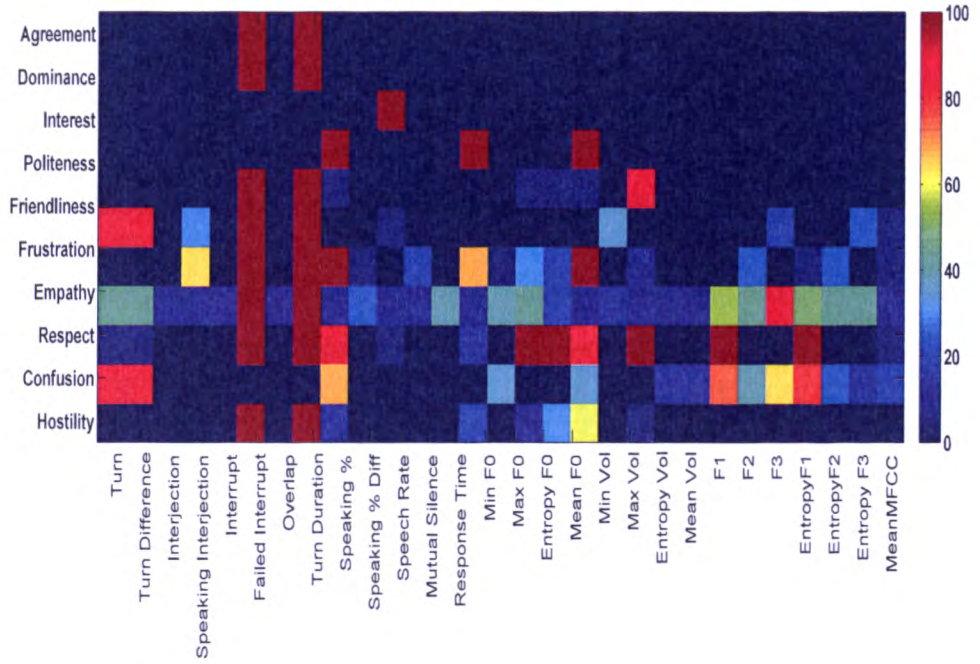


Figure 2.11: Colormap of social indicators against the audio features for AC.

Section 2.4 and the colormap in Fig. 2.11. For example, *Agreement* is strongly and moderately negatively correlated with Interrupt and Overlap respectively, and both these features occur most frequently in the selected feature set that obtains the best accuracy. Similar arguments can be drawn for *Interest* (Appendix .2), where the social indicator is strongly or moderately related (positive or negative) to features Turn Duration, Response Time, and Mean F0. These features prove to be the most reliable set of features to obtain the best classification results. The indicator *Dominance* is related to several features such as Turn, Mean F0, or Response Time apart from Speaking Percentage Difference, but the Speaking Percentage Difference alone forms the most reliable feature in obtaining the best classification accuracy. It can be seen from the "wheel" diagrams that at least one of the features which is strongly correlated (positively or negatively) to an indicator, also occurs most frequently in the selected feature set for that indicator (see Fig. 2.11).

2.5.2.2 Multi-class Classification

The sociometrics can take three values in the AC (1-3). We trained multi-class classifiers in a supervised manner, where the (rounded) average score provided by the judges served as labels for supervised learning. In this work, we considered eight kinds of multi-class classifiers: Support Vector Machine (SVM), Support Vector Ordinal Regression (SVOR), Artificial Neural Network (ANN), k-Nearest Neighbor, Naive Bayes, Adaptive Boosting, Bagging, Random Subspace Ensembles, and Least squares Boosting, where the last four are ensemble classification methods [137,138]. The classification performance is computed by leave-one-person-out cross-validation; for each participant the classifier was tested on the instances of that participant and trained on the instances of the other participants; next the average performance across all participants was computed. We tested both linear and RBF kernels for the SVM classifier, with parameters C for linear, C and σ for RBF. These parameters were optimized using cross-validation on training part of the data. We used the parameter values which provided the best results. Similarly for KNN the number of nearest neighbors was chosen using cross-validation. For all other classifiers we used default values from *Matlab* documentation and *Matlab* toolbox [139,140]. Table 2.7 summarizes the classification results obtained for each sociometric. Table 2.8 lists the Root Mean Square error (RMSE) of the classifier with the best result along with the RMSE of a trivial classifier that always has the value 2 (medium) as output, which serves as a baseline for our assessment. The RMSE is computed between the average annotation value and the classifier output. We define the normalized RMSE (NRMSE) as follows:

$$\text{NRMSE} = \frac{\text{RMSE}}{\text{RMSE(baseline)}} \times 100. \tag{2.4}$$

This measure allows us to easily compare the performance of the different classifiers with the baseline. The values of the NRMSE for the best classifier for each social indicator (see Table 2.7) are listed in Table 2.8. A lower value of NRMSE implies that our classifier is better in detecting the social indicator than the baseline.

Table 2.7: The classification results achieved for each sociometric using various machine learning algorithms.

Sociometrics	SVM	SVOR	ANN	KNN	Naive	Bagging	LSboost	Adaboost	Subspace
Agreement	83%	76%	76%	82%	78%	**84%**	72%	64%	66%
Dominance	**86%**	86%	80%	78%	81%	79%	84%	77%	83%
Interest	82%	82%	78%	79%	74%	**85%**	82%	62%	74%
Politeness	**81%**	70%	74%	68%	71%	72%	67%	74%	74%
Friendliness	49%	42%	47%	50%	**51%**	49%	48%	48%	49%
Frustration	44%	47%	48%	46%	49%	48%	42%	48%	**50%**
Empathy	52%	**59%**	51%	42%	48%	50%	45%	52%	53%
Respect	53%	55%	49%	40%	51%	49%	46%	**59%**	52%
Confusion	**81%**	74%	77%	63%	67%	78%	75%	73%	78%
Hostility	**77%**	76%	66%	66%	70%	74%	65%	76%	76%

Table 2.8: The RMSE values for the baseline and the best classifier for each social indicator (see Table 2.7), in addition to the NRMSE values.

Social indicator	RMSE	RMSE (baseline)	NRMSE
Agreement	0.45	1.65	27%
Dominance	0.32	1.52	21%
Interest	0.38	1.53	24%
Politeness	0.49	0.86	56%
Friendliness	0.67	0.73	91%
Frustration	0.71	0.77	92%
Empathy	0.61	0.65	93%
Respect	0.59	0.75	78%
Confusion	0.43	0.87	49%
Hostility	0.49	0.91	53%

The numerical results in Table 2.7 show that *Agreement*, *Dominance*, *Interest*, *Politeness*, *Confusion*, and *Hostility* can be detected accurately from non-verbal audio features. On the other hand the accuracy for *Friendliness*, *Frustration*, *Empathy*, and *Respect* is low. Therefore, we conclude that for some social indicators non-verbal cues may not have sufficient information. As expected, classifiers with poor classification performance (see Table 2.7) have large NRMSE values (see Table 2.8).

2.6 Audio-Visual Corpus (AVC)

In this section, we discuss our analysis of the AVC. We consider three feature sets (audio, video, and audio-visual based) for all the sociometric indicators. First we elaborate on the most relevant visual features for the sociometrics. Next we explain how we infer the sociometrics from the most relevant audio and visual cues, and present classification results for the ten sociometrics.

2.6.1 Feature Selection

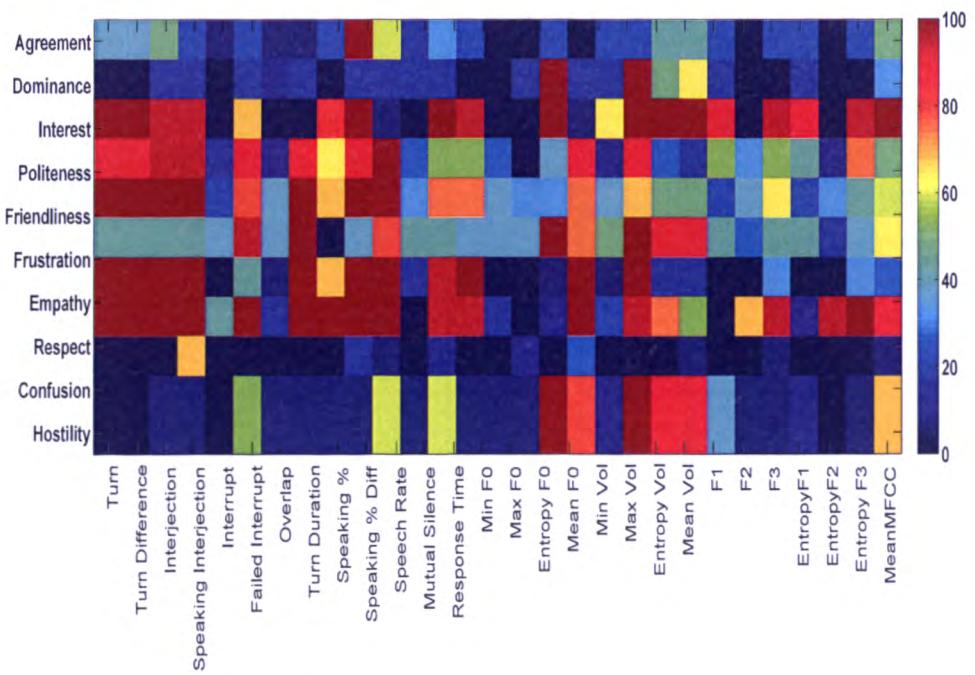


Figure 2.12: Colormap of social indicators against the audio features for AVC.

Similar to Section 2.5.2.1, we applied correlation coefficient technique [135] to determine the most salient features for inferring the sociometrics. We applied forward selection [136] to determine the best configuration of features. In Fig. 2.12 and Fig. 2.13 we show the plots of social indicators against the audio and video features respectively. We determine the best feature set that obtains the highest classification accuracy through forward feature selection. The values in the colormaps show

the frequency of occurrence of each feature in the selected feature set that obtains best accuracy for the corresponding social indicator. Therefore, we can get an idea about the significance of each feature in determining the best accuracy for the social indicator from Fig. 2.12 and Fig. 2.13.

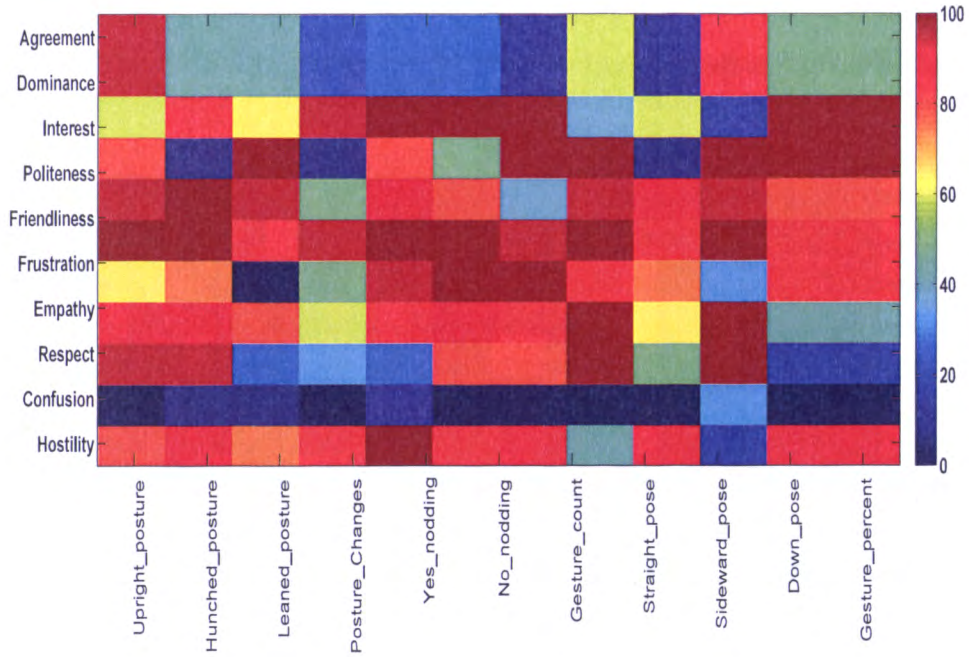


Figure 2.13: Colormap of social indicators against the video features for AVC.

The values in Fig. 2.12 and Fig. 2.13 can similarly be compared to the "wheel" diagrams from Section 2.4 and Appendix .2. The "wheel" diagrams show the correlation of features for each social indicator, whereas colormaps show the significance of each feature in obtaining highest accuracy. The comparison shows that in most of the cases at least one of the features with high value in Fig. 2.12 or Fig. 2.13 also has a strong correlation (positive or negative) in "wheel" diagrams from Section 2.4 and Appendix .2.

Table 2.9: The best classification results achieved for each sociometric. Each row shows the social indicator, where the columns represent the feature set used and the best algorithm for each indicator.

Sociometrics	Audio	Classifier	Video	Classifier	Audio-Visual	Classifier
Agreement	80%	SVM	72%	SVM	**81%**	Naive
Dominance	86%	SVOR	88%	Adaboost	**90%**	SVOR
Interest	89%	SVM	86%	ANN	**92%**	SVM
Politeness	75%	SVM	71%	SVM	**76%**	SVM
Friendliness	54%	SVOR	63%	SVOR	**63%**	SVM
Frustration	67%	SVOR	**69%**	SVM	67%	SVOR
Empathy	53%	Bagging	63%	SVM	**67%**	SVOR
Respect	54%	SVOR	60%	SVM	**62%**	SVM
Confusion	89%	SVM	88%	SVM	**89%**	SVM
Hostility	72%	SVM	**74%**	SVM	72%	SVM

Table 2.10: The RMSE values for audio-visual classification and the baseline, in addition to the NRMSE values.

Sociometrics	RMSE	RMSE (baseline)	NRMSE
Agreement	0.41	0.85	48%
Dominance	0.35	0.97	36%
Interest	0.32	0.97	32%
Politeness	0.61	0.71	85%
Friendliness	0.59	0.77	76%
Frustration	0.57	0.89	64%
Empathy	0.55	0.72	76%
Respect	0.66	0.75	88%
Confusion	0.31	0.92	33%
Hostility	0.49	0.89	55%

2.6.2 Multi-class Classification

As in the AC, the social indicators in the AVC can take three values on the likert scale (1-3). Similar to AC analysis, we trained multi-class classifiers in a supervised manner from the newly collected AVC. The (rounded) average score provided by the judges served as labels for supervised learning. We utilized the same machine learning algorithms as for the AC (see Section 2.5), and the classification performance was again computed by leave-one-person-out cross-validation.

The numerical results in Table 2.9 suggest that the audio modality crucial for infer-

ring the social indicators, and generally leads to better results when compared to the visual modality. In addition, combining the audio modality with the visual modality significantly enhances the classification. The improvement in accuracy by including visual information is 14% for *Empathy*, 9% for *Friendliness*, 9% for *Respect*, 4% for *Dominance*, 3% for *Interest*, 1% for *Agreement*, and 1% for *Politeness*. For the AVC the accuracies follow a trend similar to the AC. The accuracies for *Agreement*, *Dominance*, *Interest*, *Politeness*, *Confusion* and *Hostility* are the highest followed by *Frustration*, *Empathy*, *Friendliness*, and *Respect*. An interesting observation from Table 2.9 is the increase in accuracy for *Friendliness*, *Empathy*, and *Respect* when visual cues are included. Therefore these results indicate the significance of visual features for the inference of various social indicators. Similar to AC, classifiers with poor classification performance (see Table 2.9) have large NRMSE values (see Table 2.10).

2.7 Combined Audio Analysis of AC and AVC

Ultimately, we aim to develop systems that can provide reliable social indicators in a large variety of scenarios and conditions, e.g., for dialogs at the workplace, in the outdoors, or at informal meetings. Obviously, our AC and AVC on the other hand are collected in a controlled environment, and therefore, the results reported so far might be optimistic. To emulate a variety of recording conditions, we also conducted an experiment where we combined the audio recordings of the AC and AVC, and trained classifiers based on this merged data set. Note that the recording conditions for both the AC and AVC are quite different: The duration of conversations in both corpora is different (2.5–3 minutes vs. 1 minute); also in the AC, both speakers were participants, whereas in the AVC, only one speaker was a participant, while the other was a facilitator. To test the robustness of our approach, we combined the audio data for both corpora, and trained a 3-class classifier on this mixed data set

Table 2.11: The accuracies for 3-class classification for the AC and AVC combined. The right column shows the best performing classifier for each indicator.

Sociometric	AC-AVC	Classifier
Agreement	78%	SVM
Dominance	81%	SVOR
Interest	74%	SVOR
Politeness	65%	Naive
Friendliness	62%	SVOR
Frustration	69%	SVOR
Empathy	58%	SVOR
Respect	51%	Adaboost
Confusion	81%	SVM
Hostility	73%	Bagging

Table 2.12: The RMSE values for combined audio classification and the baseline, in addition to the NRMSE values.

Sociometrics	RMSE	RMSE (baseline)	NRMSE
Agreement	0.56	0.65	86%
Dominance	0.52	0.77	68%
Interest	0.48	0.53	67%
Politeness	0.60	0.79	76%
Friendliness	0.65	0.74	87%
Frustration	0.67	0.82	81%
Empathy	0.60	0.68	88%
Respect	0.63	0.74	84%
Confusion	0.36	0.89	41%
Hostility	0.49	0.89	55%

to assess the classification of all the social indicators. We applied the same classification algorithms as before. We performed a leave-one-person-out cross-validation analysis by means of the associated audio annotations (from the AC and AVC). To our knowledge, such analysis of mixed data, stemming from multiple independent recordings, has never been conducted in the context of automated behavioral analysis, although it is critical for assessing the robustness of the sociometrics. We also normalized the feature values by duration, to cater for different conversation

duration in both corpora.

Considering the substantial difference between both corpora, results in Table 2.11 show that we can still predict the sociometrics with reasonable reliability. The 3-class classification performance is above 70% for *Interest*, *Dominance*, *Agreement*, *Confusion* and *Hostility*. The NRMSE values from Table 2.12 confirm that classification of social indicators is better than baseline in all the cases. The difference between the classifier RMSE and the baseline RMSE is larger for social indicators with higher accuracy, except for *Agreement* where the classifier RMSE is only modestly below baseline RMSE. Overall this suggests that the system might be applied in real-life scenarios, where the context and conditions of conversations may vary significantly. Of course, additional tests in real-life scenarios would be required to further validate this claim.

2.8 Computational Complexity

In this section we provide information regarding the computational complexity of our system. For this purpose, we analyzed conversations of different durations and determined the time required for the analysis. Sociometric analysis has four major computational parts: speech detection, initialization, feature extraction, and machine learning. In speech detection, the speech and non-speech portions in the audio recordings are separated. During initialization, all variables are initialized before analysis. In the feature extraction step, the most salient prosodic and conversational features are computed. In the machine learning step, the best performing machine learning algorithms are applied to the feature sets in order to infer speech mannerisms and social indicators. In Table 2.13 we show the conversation duration in seconds in the left column, and the corresponding elapsed time for speech detection, initialization, feature extraction, and sociometric analysis in the right columns.

Table 2.13: Computation time (in seconds) for the sociometric system. From left to right: duration of the conversation (in seconds), speech detection, initialization, feature extraction, and machine learning.

Duration	Speech detection	Initialization	Feature extraction	Machine learning	Total
30	2.3	0.57	1.3	0.03	4.14
60	4.2	0.57	2.5	0.03	7.25
90	7.4	0.58	4.4	0.03	12.40
120	9.8	0.56	6.2	0.03	16.60
180	12.4	0.57	8.2	0.03	21.22
270	22.2	0.57	16.3	0.03	39.10

It is noteworthy that the time for initialization is independent of the duration of the conversation, since the number of variables is fixed. Likewise the computation time for machine learning does not grow with the length of the conversation, as the number of features is fixed. On the other hand, the time required for speech detection and feature extraction scales approximately linearly with the conversation length. As a result, also the total analysis time scales linearly with the duration of the conversation, yet remains small compared to the duration of the conversation (see Fig. 2.14). Consequently, the proposed sociometric system can be used for real-time analysis of ongoing dyadic conversations. In the next section, we will describe our implementations on several ubiquitous platforms.

2.9 Summary

Social interactions play a significant role in learning and shaping the norms and productivity in individuals, teams, departments and organizations. Recent advancement in technology has led to growing expectations of building real-time systems that can promote behavioral changes. Although increasing effort to develop systems that are capable of automatically analyzing human behavior are ongoing, numerous challenges such as decoding social interactions, providing feedback, and implementation concerns still remain. In this work we have tried to address few of these

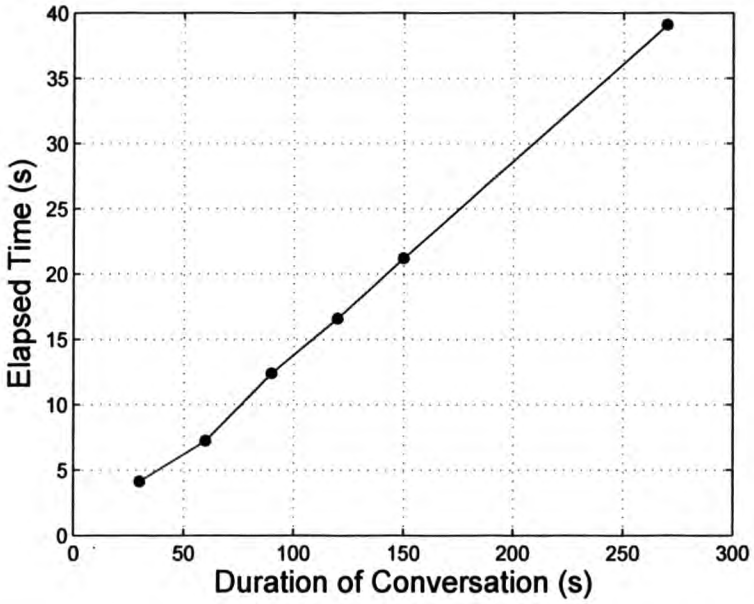


Figure 2.14: Analysis time for conversations of different durations.

issues by introducing a system capable of inferring multiple social indicators using non-verbal audio-visual features, with low computational complexity.

Chapter 3

Feedback Platforms

In this chapter, we will focus on the details of various feedback platforms that we have prepared for the Sociofeedback system. In section 3.1, we present the implementation of the real-time sociometric feedback system on VoIP (Skype) platform. In section 3.2 we present the implementation on the Android smartphone and smartglass platforms.

3.1 VoIP

Currently, VoIP (Voice over Internet Protocol) technology is ubiquitous especially because of the proliferation of such applications as *Skype*, *Viber*, and *GoToMeeting* [141–143]. To integrate our social analysis with VoIP we chose the service *Skype* [144] because of its popularity. The first challenge in this task was to be able to record each speaker on separate files. We used appropriate settings on a *Skype* recorder (a software called *SuperTintin* [144, 145] for *Skype*) to capture the audio-video data of each participant of the *Skype* call on two separate files. Subsequently, we separated the audio from the video data and fused the audio data on a single 2-channel audio file at a sampling frequency of 8 kHz, where each channel contained the audio data of only one person. The recordings were automatically stored in segments of 1 minute.

We interfaced *Matlab* [139] with the recorded audio through a file event handler which began the analysis as soon as a new recording file was saved.

In *Matlab*, the recorded files were processed as described in the previous sections,

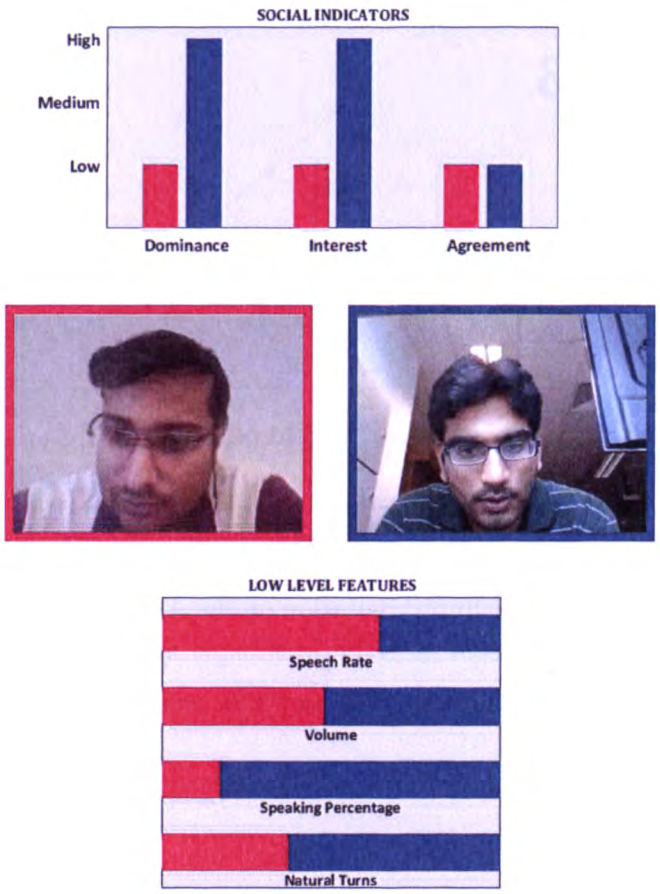


Figure 3.1: Sociometric analysis during a conversation on VoIP (*Skype*). The two speakers are shown in the bottom left, where the two audio-video streams captured from *Skype*. Non-verbal speech cues are reported in the bottom right corner, whereas the social indicators are shown at the top.

including speech detection, feature extraction, and sociometric analysis. Feedback messages may be communicated to the users through the *Skype* API [146] which allows external applications to send messages over *Skype*.

Figure 3.2: Sociofeedback on smartphones and smartglasses. Based on the social indicators, feedback messages are generated, which in turn are sent to the smartglasses(bottom left and middle) and smartphone(bottom right).

3.2 Smartphones and Smartglasses

In the VoIP implementation discussed in the previous section, the non-verbal cues and social indicators are displayed on the screen and updated continuously during the conversation. Here we first discuss an alternative approach, implemented on

the *Android* platform [147]; instead of displaying the indicators, we send messages to the user whenever the non-verbal cues and social indicators are in an abnormal range (see Table 3.1), and then introduce the smart glasses used.

3.2.1 Technical Details

Since *Android* devices often have small displays (e.g., smartglasses or smartwatches), only limited information can effectively be displayed; therefore, it is more adequate to send short messages to the user whenever feedback to the user is required on his/her speech mannerisms or social behavior, instead of displaying a plethora of measures. Although the design shown in Fig. 3.5 could easily be implemented on *Android* devices, it would be less effective for devices with small displays such as smartglasses or smartwatches. To develop a prototype application we selected *Android* as our platform of choice because if its popularity and free access. As discussed in the previous section the analysis was conducted in *Matlab* environment; we used machine learning algorithms to infer various social indicators. The challenge in this case was to interface *Matlab* with an *Android* application such that the sociometric inference could be communicated to the user on his/her *Android* device. This interfacing was achieved through the TCP server and client approach, where the TCP server operated on the same computer where the sociometric analysis was performed in *Matlab*, and listened to the incoming client connection, specifically, the *Android* device. Once the application is launched on the *Android* device, it connects to the TCP server and information from the sociometric system is sent to the *Android* device (see Fig. 3.2). In this implementation, we provide feedback to the user about the speech mannerisms and social behavior. An example of the feedback message on the Sociofeedback *Android* application can be seen in Table 3.1. Table 3.2 shows the combinations of audio features and values of the social indicators of agreement, dominance, and interest that trigger the feedback messages shown in Table 3.1.

Table 3.1: Feedback messages on *Android* smartphone (left) and smartglasses (right).

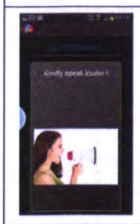
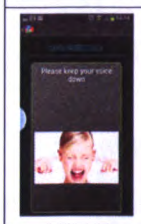
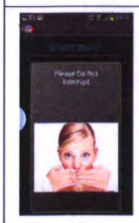
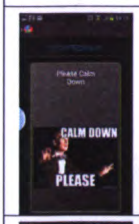
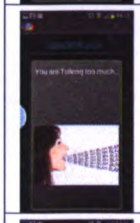
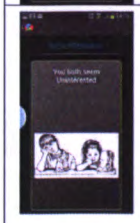
Android	Smart Glass	Description
		**Speak louder**: When a speaker is speaking too softly, the system will ask the speaker to speak at a louder voice.
		**Too noisy**: When a speaker is speaking too loud, the system will recommend the speaker to lower his/her voice.
		**Stop interrupting**: The system will ask the speaker to stop interrupting if the speaker is interrupting too often.
		**Calm down**: The system will ask the speaker to calm down if the speaker is too aggressive.
		**Slow down**: The system will ask the speaker to slow down when he/she is speaking too much.
		**Uninterested**: The system will invite the speakers to contribute more to the discussion, when both of the speakers have not been speaking for a period of time.

Table 3.2: Combinations of audio features and machine learning output that trigger corresponding feedback messages(first column).

Feedback	Speaking percentage	Speech rate	Volume	Agreement	Dominance	Interest
Speak louder			Low			
Too noisy			High			
Stop interrupting				Low		
Calm down	High			Low	High	High
Slow down		High				
Uninterested	Low					Low

3.2.2 Smartglasses

We also developed applications for two smartglasses: *Vuzix M100* and *Google Glass* [148, 149].

3.2.2.1 Vuzix smart glass

Figure 3.3: The Vuzix M100 Smart Glass.

The Vuzix M100 Smart Glass as shown in the figure below, is the first smart glass in the market. This Vuzix M100 Glass is an Android-based wearable computer, based on the Android 4.0.4 operating system, with an enhanced monocular display, speaker and a high-definition camera. It can be used as a stand-alone device and has various integration means such as, pairing with an Android device or wirelessly connecting with other devices and the Internet. In addition to the multiple choices

of connections with devices, the Vuzix M100 Smart Glass is more industrial oriented whereby the glass can also be mounted on the safety glass and headbands for operations use. For this sociofeedback interface, virtual keyboard is required to enter the IP address for client server connections. This is achieved by downloading the available Smart Glass Manager application for Android devices that provides more traditional tap and drag interface style and provides keyboard input. As such, the applications installed on the M100 Glass can be controlled and managed through the partner device via Bluetooth connection as shown in the figure below.

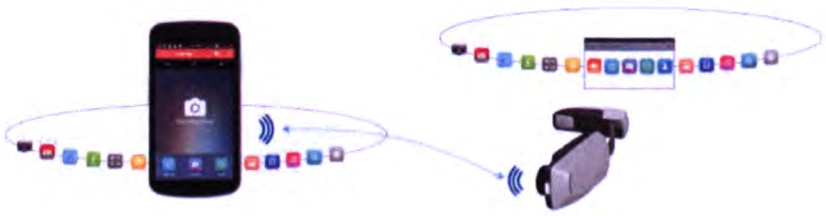


Figure 3.4: The Smart Glass Manager and the Carousel on the M100.

3.2.2.2 Google Glass

Google Glass is another Android-based headset that a user can wear just like an ordinary pair of glasses with the ability to interact using natural language commands. Google Glass uses their signature OK Glass prompt to perform a wide array of functions that smart phones and other hand held devices perform today. These include making and receiving calls, taking and sharing picture on the go and the most highly publicized function, the hands free Google navigation. Google Glass in fact goes beyond functionality of smart phones as it can be used in wet weather conditions, including rain or snow.

3.2.2.3 Smartglasses Implementation

Both the platforms use *Android* as Operating System but due to separate APIs we needed to develop separate applications for each type of smartglasses. The app

Figure 3.5: The Google Glass.

development for the *Vuzix M100* smartglasses is done in *Eclipse* with *Android Software Development Kit (SDK)* as the development environment. The application for *Google Glass* was developed in the *Android Studio 1.0* with the *Glass Development Kit (GDK)*. We applied the TCP server-client format for interfacing, as mentioned earlier. The processed audio from *Matlab* is received as a string and this string is then decoded for corresponding feedback type chosen earlier. An example of feedback messages on the Sociofeedback application on *Vuzix* and *Google Glass* can be seen in Table 3.1.

3.3 Summary

We presented a few implementations, including VoIP (*Skype*), *Android* smartphones, and smartglasses (*Vuzix M100* and *Google Glass*) which can be considered as the trending technologies of the present, and the technologies that have a significant growth potential in the future. Our system works in continuous time for these implementations as we apply a sliding window. This is different from most studies, which are done on one existing corpora in an offline manner. The implementation of our proposed sociofeedback technology on VoIP, smartglasses, and smartphones can add tremendous value to existing technologies, and also pave way to novel innovative technologies. For examples: apps on smartphones could make use of the

sociofeedback technology to infer internal states of users to promote self-awareness, positive psychology, and mindfulness [150]. Such apps would generate big data on social and emotional indicators that is valuable for understanding human behavior. Such apps can also be utilized by telemedicine (VoIP enabled) to monitor the effect of treatments or drugs on the mood of patients.

Page intentionally left blank

Chapter 4

Sociofeedback via Nao Robot

In this chapter, we present an application of the Sociofeedback system for social robotics. In section 4.1 we overview the sociofeedback system and explain how we interfaced the sociofeedback system with a Nao humanoid robot. In section 4.2 we present the Godspeed questionnaire frequently used for research in the domain of human robot interaction. In section 4.3, we present the motivation behind the two user studies that we conducted with Nao robot. In sections 4.4 and 4.5, we present the experimental design and results for the user studies and discuss our findings.

4.1 The Sociofeedback System Overview

In this section we briefly describe the sociofeedback system. The overall system is illustrated in Fig. 1.1. This system is based on our earlier work [41], where we developed a machine learning system that is able to infer the levels of interest, dominance, and agreement with 85%, 86% and 82% accuracy respectively. In the following, we first explain the hardware setup for audio recording of conversations. Next, we briefly describe the extraction of nonverbal speech cues. Then, we explain how we infer social states from those cues. Finally, we explain how Nao determines social states to provide real-time sociofeedback.

4.1.1 Sensing and Recording

We adopted easy-to-use portable equipment for recording conversations; it consisted of lapel microphones for each of the two speakers and an audio H4N recorder that allowed multiple microphones to be interfaced with the laptop. The audio data was recorded in brief consecutive segments as a 2-channel audio .wav file.

4.1.2 Extraction of Non-Verbal Cues

We considered two types of low-level speech metrics: conversational and prosody related cues. The conversational cues account for *who* is speaking, *when* and *how* much, while the prosodic cues quantify *how* people talk during their conversations. We computed the following conversational cues: the number of natural turns, speaking percentage, mutual silence percentage, turn duration, natural interjections, speaking interjections, interruptions, failed interruptions, speaking rate and response time [41]. Fig. 4.1 illustrates different conversational cues. Black indicates the time instants when the participant is speaking, whereas white depicts the periods when the participant is silent.

We considered the following prosodic cues: amplitude, larynx frequency (F0), formants (F1, F2, F3), and mel-frequency cepstral coefficients (MFCCs). These cues are extracted from 30ms segments at a fixed interval of 10ms; they tend to fluctuate rapidly in time. Therefore, we compute various statistics of those cues over a time period of several seconds, including minimum, maximum, mean and entropy to infer speaking mannerism.

4.1.3 Social State Estimation

In our earlier work [41], an audio corpus of 150 conversations the subjects from both corpora are students of Nanyang Technological University (NTU). The total number of individuals that participated in the corpus was 22, of which 17 were males,

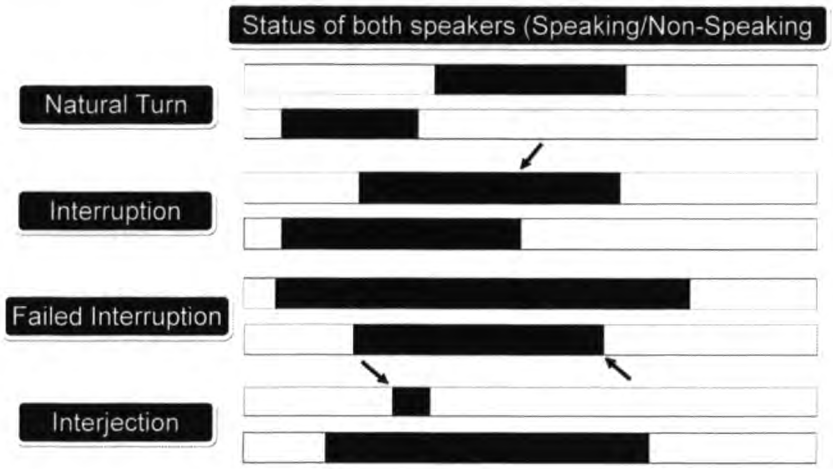


Figure 4.1: Illustration of turn-taking, interruption, failed interruption, and interjection. Those conversational cues are derived from the binary speaking status (speaking vs. non-speaking).

and 5 were females. The age of the students varies from 18 to 30. The topics of conversations ranged from discussion of assignments, projects of students, to social and political views. In some of the dialogs, there are problematic situations such as conflicts and disagreements, periods of boredom, aggressive behavior, or poorly delivered speech (e.g., low volume or fast pace).

Each recording in the corpus was annotated by multiple people ("judges"), each assessing a subset of the corpus. For each recording in the corpus, the judges completed a questionnaire related to speaking mannerisms and behavioral aspects of each participant. For example, if a participant seemed bored to the annotator, the latter would assess the interest level as "low"; in contrast, if the participant seemed excited, the annotator would quantify the interest level as "high".

The annotation served as labels for supervised learning of machine learning classifiers. Once the speech cues were calculated, they were fed to machine learning algorithms. We considered four kinds of multi-class classifiers for inferring the social state of the participants: K-Nearest Neighbor (KNN) [151], Artificial Neural Network (ANN) [152], Naive Bayes [152], and Support Vector Machine (SVM) [153]. Speaking mannerism are quantitatively assessed by low-level speech metrics of vol-

ume, and speech rate. The social behavior is quantified by sociometrics including level of interest, agreement, and dominance. Together, they provide a comprehensive picture of the social state of participants in dialogs.

We performed these calculations in Matlab on a 2GHz dual-core processor with 2GB RAM. It took approximately 3-5 seconds to perform speech detection and compute speech cues from 1 min dialogs, and to perform multi-class classification, yielding the levels of interest, agreement, and dominance. Therefore, on that computer platform, the total time required for inferring those social indicators from a 1 min dialog is about 3-5 seconds, allowing us to perform such analysis in real-time settings with limited delay. We conducted a user study of this sociofeedback system [41], by integrating it with a humanoid Nao robot.

4.1.4 Feedback Generation via Nao

As mentioned earlier, the social indicators are computed in Matlab. We integrate Nao into this system by transmitting the output of the Matlab script to Nao through the TCP/IP server-client framework. More precisely, once the Matlab script determines the social state, it sends a feedback message to Nao via TCP/IP, and Nao in turn delivers the message to the speaker(s) via speech supported by gestures. The Nao robot has 25 degrees of freedom, since it is equipped with numerous sensors and actuators, including inertial sensors, infrared and sonar receivers coupled with its axes. This multitude of sensors and actuators provide the robot with high level of stability and fluidity in its movements. However, in our experiments we only used very basic movements to simulate gestures along with text to speech generation in order to deliver the audio message. The time taken by Nao to deliver the audio message, along with gestures was approximately 3 to 4 seconds. Table 4.3 provides an overview of the feedback messages considered in this study. We chose these particular scenarios because the sociofeedback system was trained on a corpus of dyadic conversation with similar scenarios.

4.2 Godspeed Questionnaire

In this section we present the Godspeed questionnaire used for getting participant's opinion regarding their interaction with the robot. In our research we used a modified version of Godspeed questionnaire.

Table 4.1: List of Godspeed categories and the criteria associated with each category.

Anthropomorphism						
Fake	1	2	3	4	5	Natural
Machinelike	1	2	3	4	5	Humanlike
Unconscious	1	2	3	4	5	Conscious
Artificial	1	2	3	4	5	Lifelike
Moving rigidly	1	2	3	4	5	Moving elegantly
Animacy						
Dead	1	2	3	4	5	Alive
Stagnant	1	2	3	4	5	Lively
Mechanical	1	2	3	4	5	Organic
Artificial	1	2	3	4	5	Lifelike
Inert	1	2	3	4	5	Interactive
Apathetic	1	2	3	4	5	Responsive
Likeability						
Dislike	1	2	3	4	5	Like
Unfriendly	1	2	3	4	5	Friendly
Unkind	1	2	3	4	5	Kind
Unpleasant	1	2	3	4	5	Pleasant
Awful	1	2	3	4	5	Nice
Perceived Intelligence						
Incompetent	1	2	3	4	5	Competent
Ignorant	1	2	3	4	5	Knowledgeable
Irresponsible	1	2	3	4	5	Responsible

Unintelligent	1	2	3	4	5	Intelligent
Foolish	1	2	3	4	5	Sensible
Perceived Safety						
Anxious	1	2	3	4	5	Relaxed
Calm	1	2	3	4	5	Agitated
Quiescent	1	2	3	4	5	Surprised

4.3 Experiments

In this section, we explain the two sets of experiments that we have conducted. In the first set of experiments (see Section 4.4), we investigate whether the participants can understand the feedback messages delivered by Nao robot. We also explore different ways to deliver the feedback messages, viz., only by audio or by audio combined with gestures. In the second set of experiments (see Section 4.5), we investigate the perception of the Nao robot when it provides feedback for dyadic conversations. In this setting, the Nao robot provides feedback to the participants after brief conversations. Table 4.2 lists the user studies that we have conducted along with the objectives of each user study.

In our experiments we we obtained feedback from the users about the Nao robot by means of Godspeed questionnaires [66]. In the Godspeed questionnaire the participants rated their perception of the robot on different criteria. The Godspeed questionnaire contains a collection of measures for evaluating a social robot, including anthropomorphism (similarity to human form), animacy (life likeness), likeability (personal likeness of the participant), perceived intelligence, and perceived safety of the robot.

Table 4.2: List of experiments conducted and their objectives.

Experiment	Objectives
Identification of Feedback Messages	1-To determine the accuracy with which the participant can identify feedback messages delivered by Nao.
	2-To compare audio and gesture modalities for feedback delivery , assessed by Godspeed questionnaires.
Integration with the Sociofeedback System	1-To determine the accuracy with which the sociofeedback system can analyze and generate feedback messages for real conversations.
	2-To investigate how the participants assess the Nao robot by means of Godspeed questionnaires.

4.4 Experiment 1: Identification of Feedback Messages

There were 20 (16 males and 4 females) participants in this first set of experiments with a mean age of 25 and SD of 2.42. All participants are NTU students. As the medium of instruction at NTU is English, all participants could easily understand the feedback messages delivered by Nao robot. The experiments were conducted in a meeting room similar to the one shown in Fig. 1.1.

Each experiment in the first set lasted about 20 minutes, and comprised of two sessions (see Fig. 4.2). First the participants were asked to identify a random sequence of eight messages that were delivered by audio only (without gestures), next the same is repeated for messages delivered by audio and supported by gestures. After each message, there was a brief break in which the participants selected the feedback message that they believed the Nao robot had just delivered. The participants

Table 4.3: Sociofeedback delivered by the Nao robot: gestures (left) and speech (right).

Gestures	Description
	**Normal**: "Good, carry on." Nao provides this feedback when a smooth conversation is going on.
	**Uninterested**: "You both seem uninterested." Nao will invite the speakers to contribute more to the discussion, when both speakers have not been speaking for a period of time.
	**Overly talkative**: "You are talking a lot". Nao will ask the speaker to slow down when he/she is speaking too much.
	**Aggressive**: "Please calm down". Nao will ask the speaker to calm down if he/she is being too aggressive.
	**Too silent**: "I am sorry, but I cannot hear you". When one or both of the speakers are speaking too softly, Nao will ask them to increase their volume.
	**Too loud**: "Please lower your volume". When the speakers are speaking too loudly, Nao will give feedback about the noise.

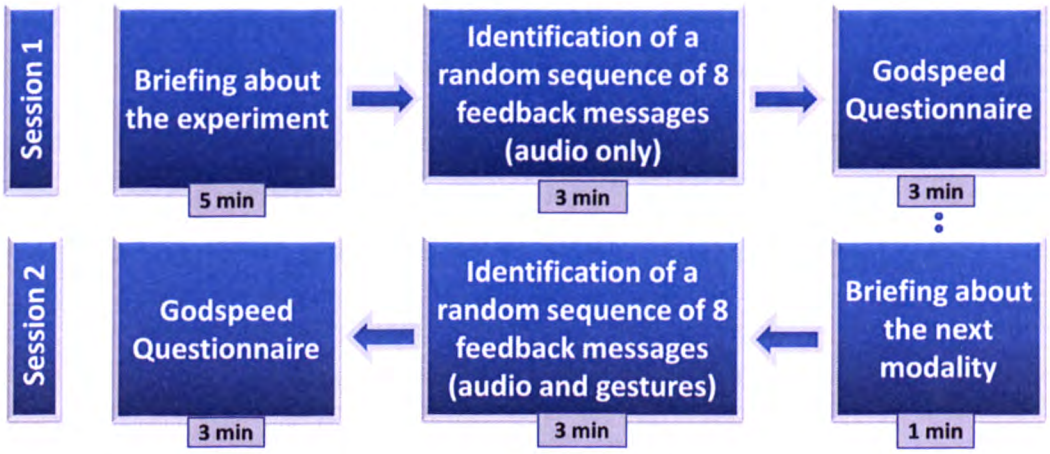


Figure 4.2: Different components of the experimental procedure. The experiments last about 20 minutes, with estimated duration of each component as indicated.

Table 4.4: Percentage of correctly identified feedback messages. Results are shown for each of the feedback messages, delivered by audio only messages and by a combination of audio and gestures.

Modality	Too silent	Too loud	Aggressive	Overly talkative	Uninterested
Audio	84.2%	68.4%	89.4%	89.4%	78.9%
Combined	84.2%	100%	94.7%	100%	100%

were given the possible answers and they had to choose one of them. After each of the two sequences of eight messages, the participants were asked to complete a Godspeed questionnaire about their experience with the Nao robot.

In this experiment, the participants were not asked to be a part of an active conversation; instead they were briefed about the context, and were then asked to judge the sociofeedback delivered by the Nao robot (feedback messages illustrated in Table 4.3). The participants were not informed about the correct answers after each session, in order to minimize the learning effect.

With these experiments, we aimed at testing the hypothesis that feedback messages can be identified more accurately when Nao uses gestures along with audio as compared to only audio feedback. We also hypothesized that the feedback delivered by

both audio and gestures will be rated higher on Godspeed questionnaire criteria, as compared to feedback provided by only audio messages.

In our questionnaire (see Table 4.5), there were two questions associated with each of the five measures of Godspeed questionnaire. Our questionnaire was a subset of the original Godspeed questionnaire, as we wanted to keep the experiments short.

4.4.1 Results for Feedback Identification

Our results are summarized in Table 4.4, showing how often (percentage) each of the feedback messages were correctly classified in each of the sessions. It can be seen from Table 4.4 that most of the feedback messages seem to be perfectly understandable when the audio messages are combined with gestures. There is room for improvement for the "Too silent" scenario. It is also clear from Table 4.4 that combining audio messages with gestures helps to improve the clarity of the feedback messages, as compared to audio messages only. To verify whether this improvement is statistically significant, we applied a repeated measures single-factor ANOVA statistical test to the responses of the 20 participants. The p-value associated with audio only vs. combined audio and gestures equals 0.027, which is clearly below 0.05, thus the corresponding improvement in accuracy is indeed statistically significant. These statistics indicate that sociofeedback is easier to identify when delivered through both audio and gestures. Also the results show that audio plays a more vital role in the delivery of the feedback messages while gestures help in improving the clarity of the message.

4.4.2 Results for the Godspeed Questionnaire

Our results are summarized in Table 4.5, showing the average scores for both conditions and the corresponding p-values of repeated measures single-factor ANOVA test. It can be seen that for each of the five measures (except perceived safety), at least one of the two questions is having a significant change in its value. Specifically,

the Godspeed scores are higher for feedback that includes both audio and gestures. The score for likeability is the highest, and the change in value for friendliness is

Table 4.5: Average values of Godspeed questionnaire(5-likert scale).

Characteristics	P-values	Average (audio)	Average (combined)
Anthropomorphism Machine/human like	**0.017**	2.94	3.52
Moving rigidly/elegantly	**0.002**	2.36	3.26
Animacy Mechanical/organic	0.129	2.89	3.26
Inert/interactive	**0.046**	3.15	3.63
Likability Dislike/like	0.110	4.26	4.63
Unfriendly/friendly	**0.008**	3.73	4.26
Perceived Intelligence Ignorant/knowledgeable	**0.009**	3.31	3.63
Unintelligent/intelligent	0.186	3.42	3.57
Perceived Safety Calm/agitated	0.741	2.36	2.26
Quiescent/surprised	1	2.84	2.84

significant. In other words, the participants seem to like the Nao robot, and by including gestures, the robot is perceived as even more friendly. Anthropomorphism also has good ratings, and the increase in values by adding gestures is significant for both questions. Moreover, the interactivity of the robot increases significantly when gestures are included. Likewise, the participants perceive the robot as more knowledgeable when it uses gestures, but the value does not change significantly for the intelligence shown by the robot. The low perceived safety values suggest that the participants were calm and quiescent in the presence of the robot, since

the minimum and maximum value correspond to calmness and agitation respectively. Interestingly, when gestures are added, the participant perceived the robot's behavior as slightly more safe (albeit a small change).

4.5 Experiment 2: Integration with the Sociofeedback System

The were 20 (17 males and 3 females) participants in the second set of experiments with a mean age of 23 and standard deviation of 2.42. The total duration for each experiment session was around 20 minutes. The aim of this second set of experiments is to investigate whether Nao can deliver feedback for a two-person dialogs. We invited participants to have a scenario-based conversation. In each scenario, the participants were asked to behave behave according to four scenarios: "normal", "uninterested", "overly talkative", and "aggressive", corresponding to the first four situations listed in Table 2. The bottom two scenarios in that Table are less interesting as social states, and hence are not considered in this experiment. In order to facilitate the scenario-based conversations, we asked the participants to follow scripted conversations. From our earlier experiments, we learned that it is difficult for participants to enact a scenario if both speakers are invited subjects. Therefore, in this user study we changed our approach such that one of the two speakers was an invited participant while the other speaker was appointed by us to serve as facilitator of the conversations. Each conversation lasted about 60 to 70s, and was analyzed in real-time by the sociofeedback system described in Section 3 (see also Fig. 1.1).

The experiment was conducted as follows (see Figure 4.3):

- First, we setup the recording system properly.
- The two speakers sat about 1.5m apart so that each microphone only recorded the voice of the respective speaker, and there was no interference from the other

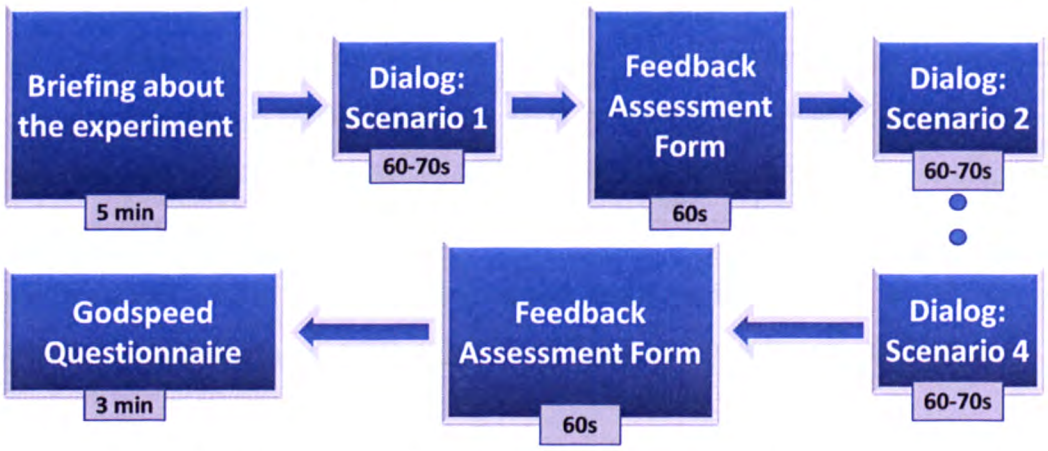


Figure 4.3: Different components of the experimental procedure. The experiments last about 20 minutes, with estimated duration of each component as indicated.

speaker.

- We attached the lapel microphones to the speakers in proper manner, in order to obtain a high-quality recording.
- The participant and the facilitator had scenario based conversations. Each conversation was about one minute in duration.
- Nao robot gave feedback after each conversation, depending on the scenario.
- The participant filled a questionnaire after each conversation, in order to rate the feedback delivered by the robot.

Table 4.6: Relationship between the social scenarios and the social indicators of interest, dominance and agreement.

Scenario	Interest	Dominance	Agreement
Normal	Medium	Medium	High
Uninterested	Low		
Overly talkative	High (Low for the other speaker)		
Aggressive	High	High	Low

Table 4.7: Percentage of correctly delivered feedback messages.

Normal	Uninterested	Overly talkative	Aggressive	Overall
100%	90%	85%	100%	93.8%

- At the end of the experiment, the participant completed the Godspeed questionnaire in order to rate the Nao robot.

4.5.1 Accuracy of Sociofeedback System

The participants were asked to act according to the first four scenarios listed in Table 2. If the feedback message delivered by Nao is in accordance with the enacted scenario, it is considered accurate. In Table 4.6 we present the relationship between the social scenarios and the values of interest, dominance and agreement. In Table 4.7 we list the accuracy of the feedback for each scenario and also present the overall accuracy of the system. In Table 4.8 we show the confusion matrix for these scenarios.

Table 4.8: Confusion matrix showing the classification results of first four scenarios. The feedback generated in these scenarios used interest, dominance and agreement sociometrics predicted by means of an SVM classifier.

	Normal	Uninterested	Overly talkative	Aggressive
Normal	20	0	0	0
Uninterested	0	18	0	2
Overly talkative	1	0	17	2
Aggressive	0	0	0	20

As seen from Table 4.7 the overall accuracy of the feedback messages is 93.8%. In the cases of "Normal" and "Aggressive" scenarios all the conversations generated correct feedback but there were mistakes in other scenarios. False detections can

occur when the participants do not strictly follow the scenario. We also asked the participants whether they agreed with the provided feedback, resulting in a average score average rating of 4.5 on a scale of 5 (see Table 4.10). This shows that the participants mostly agreed that the feedback provided by the Nao robot was appropriate for the scenario.

4.5.2 Assessment of Sociofeedback via Nao

At the end of each conversation, the participants were asked to complete an assessment about the received feedback message. The questions concern different aspects of the feedback, including the content of feedback, likability, and timing (see Table 4.9). At the end of all the conversations, the participant rated his/her experience of Nao via a Godspeed questionnaire. Table 4.11 shows the average ratings for each of the Godspeed criteria. In order to keep the assessments consistent, we adopted a 5-likert scale for both questionnaires.

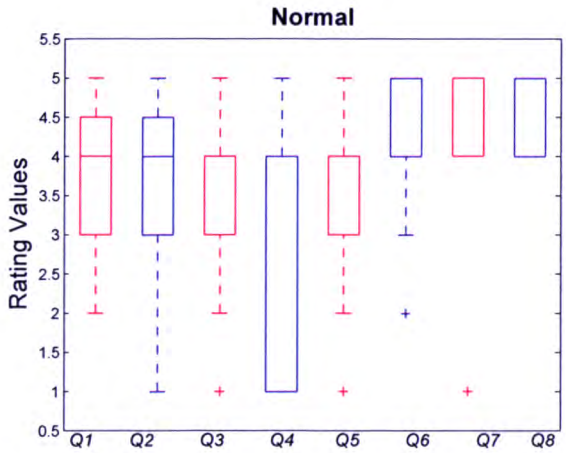
Fig. 4.4 and Fig. 4.5 display the eight ratings for each of the feedback messages. As can be seen from these figures, the ratings are mostly high. The average ratings for each question ($Q1 - Q8$) can be seen in Table 4.10. $Q1$ and $Q2$ asked the participants if they could tell when Nao was addressing them or the other speaker. The high values for all the cases implies that the participants were able to distinguish among feedback messages meant for them and the other speaker. In $Q3$, we asked participants about the timing of the feedback. Although most participants stated that Nao gave feedback timely, there is still room for improvement. The ratings of $Q4$ suggests that participants at times felt that they were interrupted by Nao. The timing can be improved by waiting for the speaker to stop his/her sentence or by getting the attention of the speaker using some gesture, before delivering the feedback message. Furthermore, the high ratings for $Q5$ and $Q6$ suggest that the interaction between Nao and the participants was fairly natural and Nao spoke with clarity. In $Q7$ we asked whether the participants agreed with the feedback message. The rating for this question is close to 5, indicating that participants agree with the

Table 4.9: Questions of the assessment form.

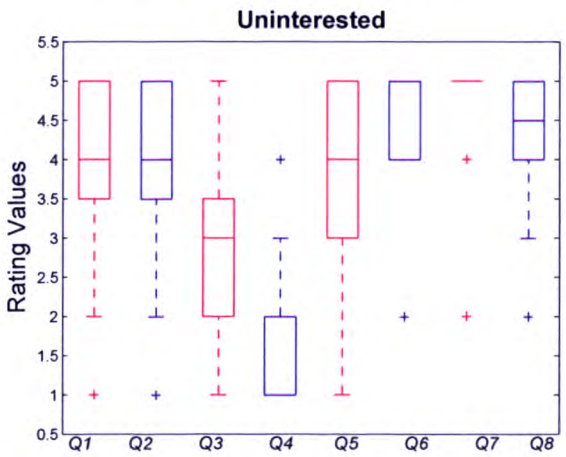
	Question
Q1	Did you notice when the sociofeedback system was addressing you?
Q2	Did you notice when the sociofeedback system was addressing others?
Q3	Was the timing of sociofeedback appropriate?
Q4	Did the sociofeedback system interrupt the conversation?
Q5	Was the interaction natural?
Q6	Did you understand the message given by the sociofeedback?
Q7	Do you agree with the given feedback?
Q8	Did you enjoy using the sociofeedback system?

feedback. Similarly, high ratings for $Q8$ confirm that participants like the feedback from Nao. Each column shows the average ratings for different scenarios.

The scores for likeability are the highest. In other words, the participants seemed to like Nao, and perceived it as friendly. Anthropomorphism also has good ratings. The robot is rated strongly human-like but the motions of the robot can be improved to make it more elegant. The animacy of Nao is also rated high by the participants, consequently, Nao was considered as highly interactive. Likewise, the participants perceived the robot as knowledgeable and intelligent. However, Nao received moderate ratings for its perceived safety, suggesting there is a room for improvement to make the participants more comfortable in the presence of Nao. Perceived safety is

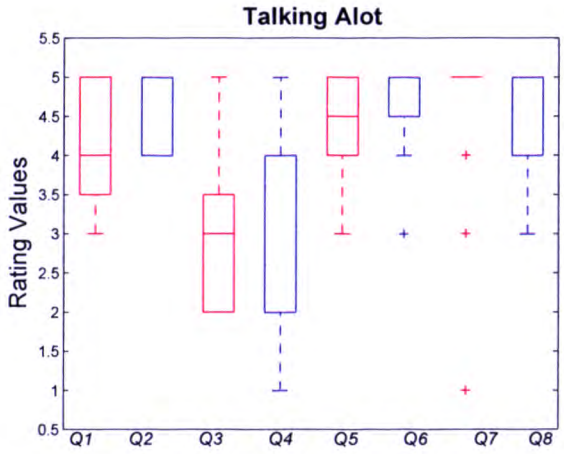


(a) Ratings for "Normal" scenario.

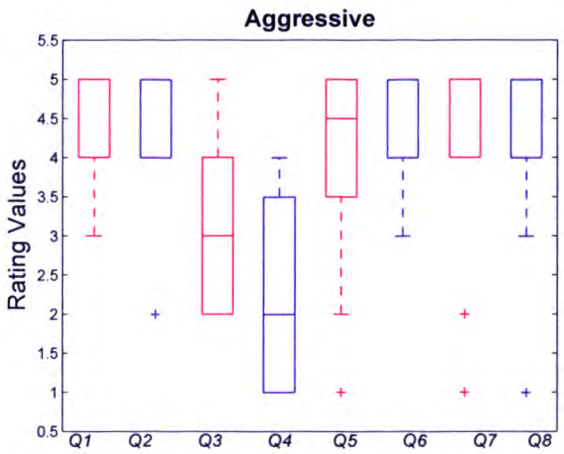


(b) Ratings for "Uninterested" scenario.

Figure 4.4: Box plots of participant's ratings for "Normal", and "Uninterested" scenarios.



(a) Ratings for "Overly talkative" scenario.



(b) Ratings for "Aggressive" scenario.

Figure 4.5: Box plots of participant's ratings for "Overly talkative", and "Aggressive" scenarios.

related to the size of the robot. Nao is a small robot (2 feet); when people interact with Nao while they are standing, the safety value is usually high [154]. In our case, Nao is seated very close to the participants (see Fig. 1.1), which may explain why the safety value is moderate in our experiments.

We also asked the participants whether they would like to receive sociofeedback or not. Out of 20 participants, 19 responded in favor of receiving sociofeedback.

At the end of the experiment, we asked to participants to leave any suggestion that they might have about the experiment. Some participants suggested improvements for the feedback messages. These suggestions were about the timing of the feedback, and also about making the feedback more natural. For instance, one participant suggested the following: "The conversation was interrupted while we were talking happily. When people are having a good conversation, it's better to use body language only instead of voice".

4.6 Summary

In summary, we can come to two conclusions based on our experiments. The first one being that people can understand social feedback messages delivered by Nao humanoid robot and they like the ability of the robot to provide social feedback.

Table 4.10: Average ratings of each assessment question. Each column shows the ratings for each question, where each row represents a social scenario.

Question	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Normal	4	4	3	2	4	5	4	5
Uninterested	4	4	3	1	4	4	5	4
Overly talkative	4	5	3	3	4	5	5	5
Aggressive	5	5	3	2	4	5	4	5
Total average	4.3	4.5	3	2	4	4.8	4.5	4.8

Table 4.11: Average ratings for the Godspeed questionnaire (5-likert scale).

Characteristics	Average Values
Anthropomorphism Machine/human like	4
Moving rigidly/elegantly	3
Animacy Mechanical/organic	4
Inert/interactive	4
Likability Dislike/like	5
Unfriendly/friendly	4
Perceived Intelligence Ignorant/knowledgeable	4
Unintelligent/intelligent	4
Perceived Safety Calm/agitated	3
Quiescent/surprised	3

Secondly the experiments validated the sociofeedback system as the accuracy for delivering appropriate feedback is really promising.

Chapter 5

Non-verbal Analysis of Schizophrenic patients' Interviews

In this chapter, we present our study that explores the relation between non-verbal speech and visual cues and negative schizophrenia symptoms. In section 5.1, we present non-verbal audio and visual cues that we have used for this study. In section 5.2, we explain the experiment design and introduce the negative symptoms (NSA-16) for which the psychologists rate the schizophrenic patients, and we also explain the cognitive remediation therapy (CRT) which the subject group is undergoing. In section 5.3, we detail the analysis done and present our findings. In section 5.4, we discuss the results achieved in this study so far.

5.1 Non-Verbal Cues

In this section we will discuss the non-verbal audio and visual features extracted from the collected data.

5.1.1 Non-Verbal Speech Cues

We computed the following conversational cues: *the number of Natural Turns*, *Speaking Percentage*, *Mutual Silence Percentage*, *Turn Duration*, *Natural Interjections*,

Table 5.1: List of conversational features considered in this study.

Category	Features
Conversational	
Speaking duration	Speaking % , mutual silence, Difference in speaking %, overlap, response time
Speaking turns	Natural turns, turn duration
Interruption	Interruptions, failed interruptions
Interjection	Interjection, speaking interjection
Prosodic	
Frequencies	Larynx frequency (F0), formant (F1, F2, F3)
MFCC	Mel-frequency cepstral coefficients
Amplitude	Mean volume, max volume, min volume, entropy

Speaking Interjections, *Interruptions*, *Failed Interruptions*, *Speaking Rate* and *Response Time* [42]. We also computed the following prosodic cues: *amplitude*, *larynx frequency (F0)*, *formants (F1, F2, F3)*, and *mel-frequency cepstral coefficients (MFCCs)*.

5.1.2 Visual Cues

We collected the video and depth data for the patient using Microsoft Kinect sensor. The depth data allowed us to extract the skeleton of the patient. Analysis of a single video takes significant time, therefore as a preliminary step we focused on just overall body movements to see if it reveals anything interesting. As this is an ongoing experiment, in future we can extract further visual features to enhance the analysis. We extracted the change in values for each joint and then calculated overall movements of these joints. The extracted joints are listed in Table 5.2. Fig. 5.1 shows the extracted skeleton for one of the patients. The patient is in seated mode which is not ideal for kinect skeleton detection but it still works pretty good as can be seen in Fig. 5.1.

Table 5.2: List of upper and lower body joints extracted from the skeleton obtained using Kinect depth data.

	Body Part	Joints
Upper Body	Elbow	Elbow left, Elbow right
	Shoulder	Shoulder left, Shoulder right, Shoulder Center
	Hand	Hand left, Hand right
	Wrist	Wrist left, Hand right
	Spine	Spine
Lower Body	Feet	Foot left, Foot right
	Hip	Hip left, Hip right, Hip Center
	Knee	Knee left, Knee right
	Ankle	Ankle left, Ankle right

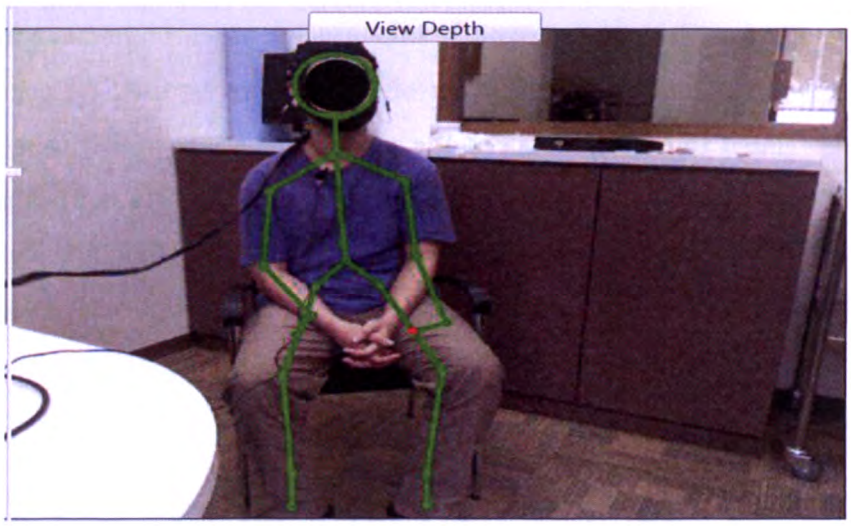


Figure 5.1: The extracted skeleton for one of the patients.

5.2 Experiment Design

We are conducting this study in collaboration with the Institute of Mental Health (IMH) in Singapore. It is an ongoing study, and the results presented here are for the cases who have completed the 12-week study. There are two groups of participants:

Table 5.3: Demographics of participants in the study.

	Subjects (N=11)		Controls (N=9)	
	Mean	Range	Mean	Range
Age	31.5	23-49	30.7	25-39
	N	%	N	%
Female	7	63.6	6	66.6
Male	4	36.3	3	33.3
Ethnicity (Chinese)	8	72.7	9	100.0
Ethnicity (Malay)	2	18.2	0	0
Ethnicity (Indian)	1	9.1	0	0
Below University Level	9	81.2	9	14.29
Above University Level	2	18.2	0	0

Subjects who are patients with schizophrenia undergoing CRT [155] in sessions at IMH, and *Controls* who are patients with schizophrenia at IMH, matched for age, gender, ethnicity, and education, but not undergoing CRT treatment. The control patients suffer from less severe cognitive impairments of schizophrenia compared to the subjects. The subject and the control groups were assessed for cognitive impairments related to schizophrenia at the start of the study period using the brief assessment of cognition in schizophrenia (BACS) tool [156]. The subject group had a mean BACS Composite score of 27.52, whereas the control group had a mean score of 42.49 on the same metric, a higher score indicating lesser cognitive impairment. The participants are recruited by IMH based on the recommendations of clinicians. The participants are provided with monetary compensation for participation in the study. The participants are all adults above 21 years of age, and have provided written informed consent. All experiments are performed in accordance to the relevant guidelines and regulations, and the study protocol has been approved by the National Healthcare Group's domain-specific Review Board in Singapore. So far we have collected data for 20 completed participants, including 11 subjects and 9 controls. Table 5.3 gives the demographics data of the participants.

The experiment has been designed such that each participant is assessed at three time-points: the first at week 0 (before the start of the CRT sessions), the second at week 2 and the third at week 12, at the completion of CRT. Each session consists of

cognitive tasks, clinical interview, and functioning tasks. During this duration of 12 weeks, each patient in the *Subject* group undergoes CRT of 2 sessions per week, so 24 sessions in total. Each CRT session lasts for about 1.5 hours; so in totality, a *Subject* is subjected to $24 \times 1.5 = 36$ hours of CRT. We will discuss the analysis of audio data acquired during the structured clinical interview. A trained psychologist from IMH conducts these interviews in English and rates each participant on a scale of 1-6, where 1 indicating no symptoms and 6 indicating severe negative symptoms, on the Negative Symptom Assessment (NSA-16) tool [157]. There is no pre-determined duration for the interview, instead it depends on participant's response to the questions asked by the psychologist. On average, the interviews lasted about 30 minutes. We have analyzed each interview in its entirety. Therefore, in this study we have analyzed about 30 hours of audio recordings (0.5 hour/interview $\times$ 3 sessions $\times$ 20 patients). In section 5.2.1 we explain the data collection procedure and in section 5.2.2 we explain the negative symptoms assessment tool used for this experiment.

5.2.1 Data Collection Procedure

We recorded the audio, video and depth data for the participant and psychologist interviews. The audio data was acquired using lapel microphones worn by both participant and the psychologist. The video and depth data was recorded only for the participant using Microsoft Kinect device. We synchronized the audio and video/depth data collection such that both software applications started with a single click.

The procedure for data collection is listed below:

1. We save the speech in a 2-channel audio .wav file (one channel for each speaker), so that we can easily detect who is speaking at any given time. We ensure that audio recording device is working properly and the participant and the psychologist wear their respective microphones.

2. We record video/depth data for each participant by means of *Microsoft Kinect* device. We make sure that participant is sitting at a reasonable distance and his/her skeleton can be detected properly.
3. Once all the devices are connected and the interviewer is ready we start the data acquisition. We monitor the software applications from another room via remote desktop to ensure proper data collection.

5.2.2 Negative Symptoms Assessment

The symptoms of schizophrenia, are frequently grouped into two general classes: positive and negative manifestations. The behaviors that are not present under normal conditions like having hallucinations and being delusional are termed as positive symptoms. Negative symptoms refer to the lack of grooming, use of language and communication abilities. A few measures or rating scales have been created to rate the positive and negative symptoms of schizophrenia. The scale used for rating negative symptoms in this experiment is shown in table 5.4.

Table 5.4: List of NSA criteria and their explanation.

Label	Criteria	Explanation
NSA 1	Prolonged time to respond	The subject pauses for inappropriately long periods before answering.
NSA 2	Restricted speech quantity	Ratings on this item suggest that the subject gives brief answers to questions and/or provides elaborating details only after the interviewer prods him.
NSA 3	Impoverished speech content	The subject may talk a lot or a little but the information conveyed is very limited.
NSA 4	Inarticulate speech	The subjects speech cannot be understood because enunciation is poor.

NSA 5	Emotion: Reduced range	Emotion is the feeling content of a persons inner life. This item assesses the range of emotion experienced by the subject during the last week (or other specified time period).
NSA 6	Affect: Reduced modulation of intensity	This item assesses the subjects modulations of intensity of affect shown during the interview while discussing matters that would be expected to elicit significantly different affective intensities in a normal person
NSA 7	Affect: Reduced display on demand	This items assesses the subjects ability to display a range of affect as expressed by changes in his/her facial expression and gestures when asked by the interviewer to show how his/her face appears when he/she feels happy, sad, proud, scared, surprised, and angry.
NSA 8	Reduced social drive	The desire of the subject to initiate social interactions. The attempts made by the subject to establish social contact with others can help measure his/her social drive.
NSA 9	Poor rapport with interviewer	This item assesses the interviewers subjective sense that he/she and the subject are actively engaged in communication with one another.

NSA 10	Interest in Emotional and Physical Intimacy	This item assesses how much the subject retains interest in emotional and physical intimacy or sexual activity.
NSA 11	Poor grooming and hygiene	The assessment of the physical grooming of the subject.
NSA 12	Reduced sense of purpose	This item assesses whether the subject possesses integrated goals for his/her life.
NSA 13	Reduced interests	This item assesses the range and intensity of the subjects interests.
NSA 14	Reduced daily activity	This item assesses the level of the subjects daily activity and his/her failure to take advantage of the opportunities his/her environment offers.
NSA 15	Reduced expressive gestures	Gestures and body movements that normally facilitate communication during speech are less than normal, or are not observed at all.
NSA 16	Slowed movements	This item assesses how much the subjects voluntary movements are slowed. At a minimum one should rate movements as gait and those of rising from a chair.
NSA 17	Global negative symptoms rating	This item assesses the overall impression of negative symptoms in the subject.
NSA 18	NSA total	Sum of the ratings from questions 1-16.
NSA 19	NSA communication	Sum of the ratings from questions 1-4.
NSA 20	NSA emotion affect	Sum of the ratings from questions 5-7.

NSA 21	NSA social involvement	Sum of the ratings from questions 8-10.
NSA 22	NSA motivation	Sum of the ratings from questions 11-14.
NSA 23	NSA retardation	Sum of the ratings from questions 15-16.

Table 5.4 lists the questions in NSA. The last six criterion namely *NSA total*, *NSA communication*, *NSA emotion affect*, *NSA social involvement*, *NSA motivation* and *NSA retardation* have composite scores. The scores for these criteria are calculated by accumulating the scores for the basic questions.

5.2.3 Cognitive Remediation Therapy

CRT is a treatment designed for cognitive rehabilitation. Research supports the effectiveness of cognitive remediation therapy for traumatic brain injury, and schizophrenia [155, 158, 159]. It helps improve cognitive abilities such as attention, working memory, cognitive flexibility and planning, and executive functioning. An improvement is these areas leads to an overall enhanced social functioning, and helps patients in obtaining and working in competitive jobs.

5.3 Analysis and Results

In this section we will present our analysis of the data collected so far. We removed prosodic features, since they were not contributing that much to the analysis. In this work we have focused on the correlations between NSA and conversational features. First, we show the correlation between the objective audio features and the subjective negative symptoms ratings, then we present our results for automated prediction of negative symptoms from audio features. At the end we present the results for the classification of the control and subject groups. The classification performance was computed by leave-one-person-out cross-validation, i.e., for each

participant the classifier was tested on the instances of that participant and trained on all the remaining instances.

5.3.1 Non-Verbal feature analysis

In this section we plot the non-verbal speech features to see the differences among subjects and control groups.

5.3.1.1 Speech features

It can be seen from Figs. 5.2, 5.3 and 5.4 that controls group has higher averages for number of natural turns, average turn duration, speaking % and speech rate, whereas the subject group has higher averages for response time and mutual silence. In summary it is clearly evident that control group participates more in the conversation as compared to the subject group.

5.3.1.2 Visual features

We extracted the skeleton of the patients from Microsoft Kinect video and depth data. The skeleton consists of various joint values that helped us determine the movement features over the course of the interview. We used the joint movement values and combined them to get three major measure i.e. lower body movement, upper body movement and an overall measure shown in Figs. 5.5 and 5.6 respectively. From the figure it can be seen that the subjects move more on average as compared to the controls.

5.3.1.3 Correlation Analysis

We extracted conversational, and visual features from the patient interviews conducted by the psychologists at IMH. Fig. 5.7, 5.8 and 5.9 show the linear correla-

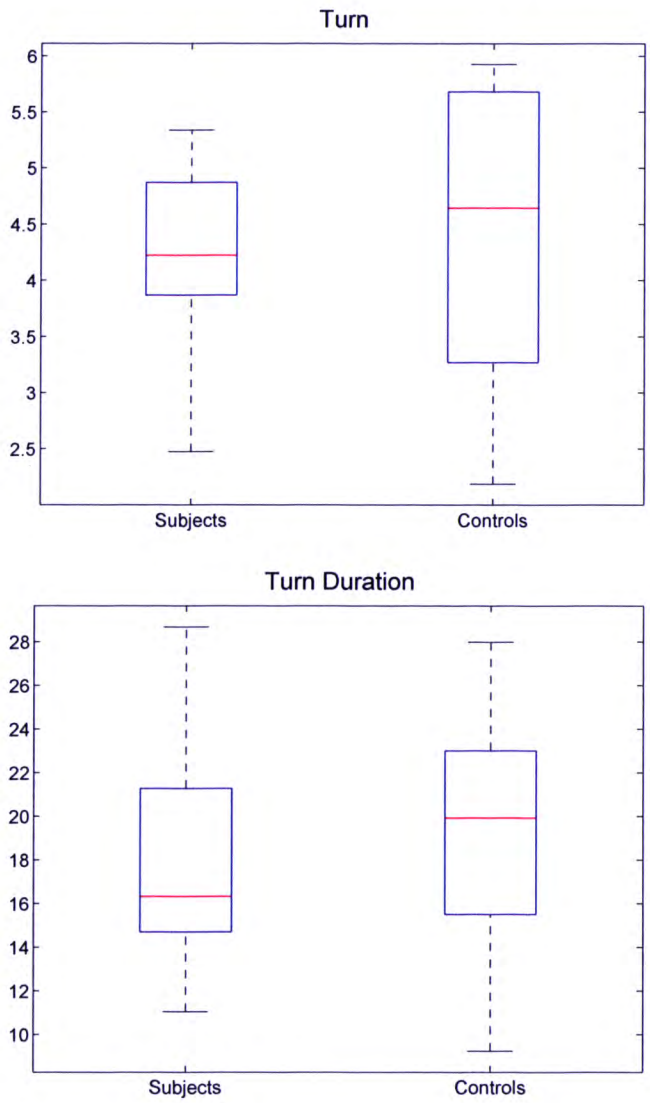


Figure 5.2: Natural turns and turn duration plots for subject and control groups.

tion of conversational speech and visual features with negative symptoms assessment done by IMH psychologists.

We calculate the linear correlation value $\rho_{X,Y}$ between a non-verbal cue x_i of the ith recording of one person, say *Turn Duration*, with the rating y_i of the same recording,

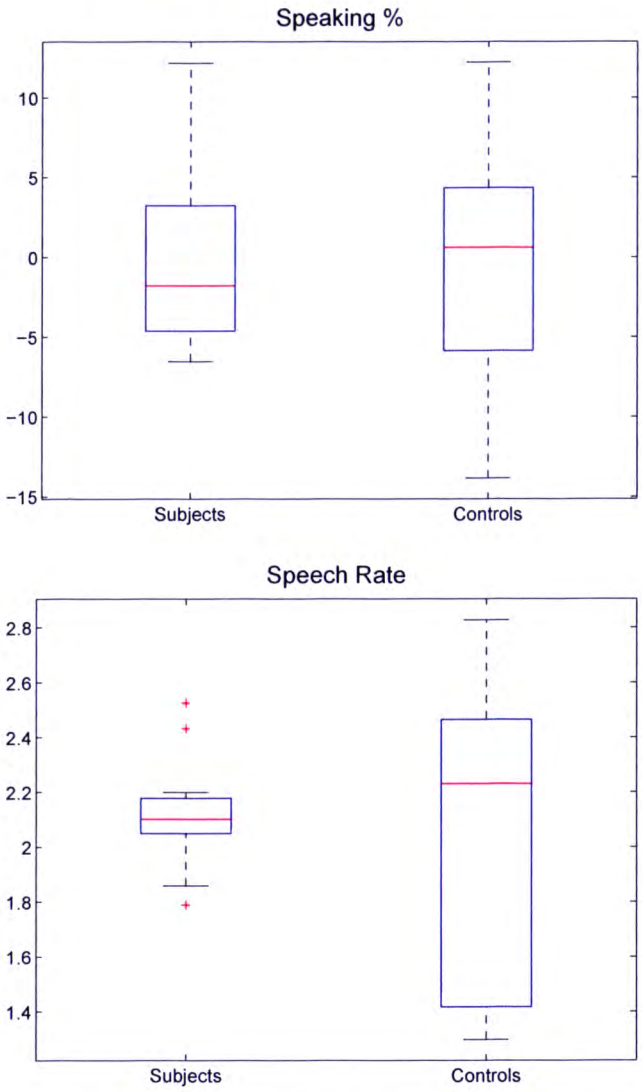


Figure 5.3: Speaking % and speech rate plots for subject and control groups.

say *Restricted Speech Quantity*, as follows:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{5.1}$$

where $\bar{x}$ and $\bar{y}$ denote the corresponding mean value for all recordings. The col-

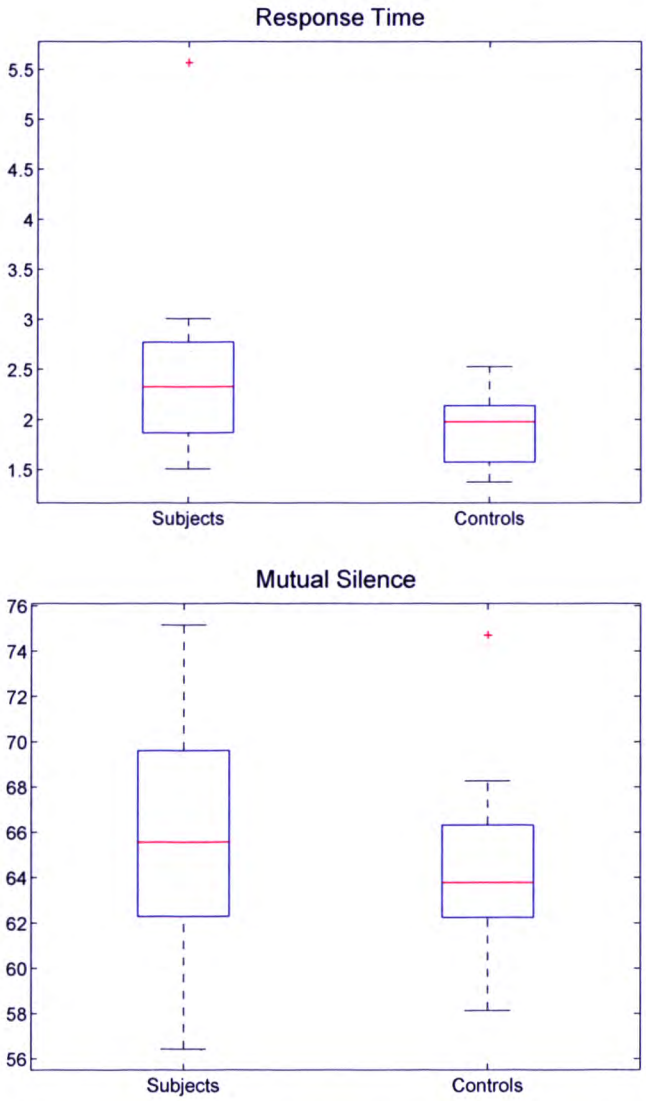


Figure 5.4: Response time and mutual silence plots for subject and control groups.

ormaps in Fig. 5.7, Fig. 5.8 and Fig. 5.9 show the correlation of NSA-16 criteria with audio-visual features. The first colormap illustrates the correlation for NSA-16 questions 1-8, the second colormap for questions 9-16 and the third colormap has the correlation values for questions 17-23. The values range from -0.7 to 0.7, with -0.7 represented as dark blue and 0.7 represented as dark red. It can be seen from the colormap that features like *Failed Interrupt*, *Overlap*, *Mutual Silence*, *Response Time*, and overall body movement are directly correlated to the negative symptoms;

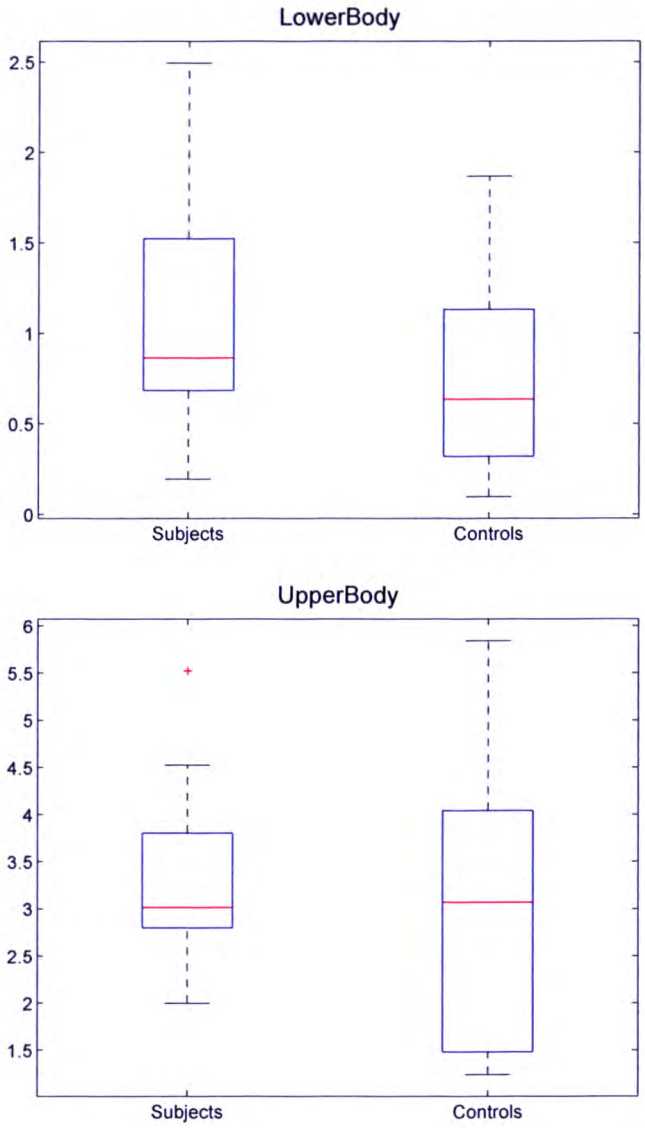


Figure 5.5: Lower body and upper body movement plots for subject and control groups.

on the other hand *Interjections*, *Interrupts*, *Speaking Percentage*, *Turn Duration*, *upper body movement*, and *overall movement* are inversely correlated to the negative symptoms.

As a next step in our analysis we determined the prediction accuracy for each NSA-16 criterion from two multi-class pattern recognition classifiers, viz., Support Vector

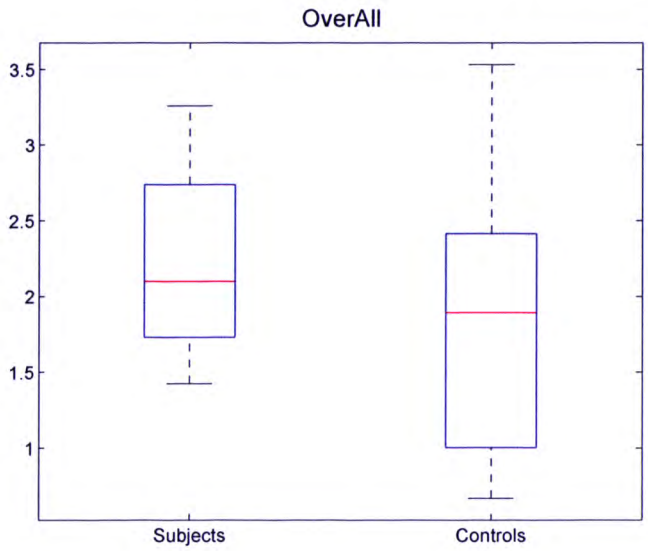


Figure 5.6: Overall movement plots for subject and control groups.

Table 5.5: Accuracies of predicting Negative Symptoms using non-verbal features.

Negative Symptoms	**SVM**	**MAE**	**RMSE**	**SVR**	**MAE**	**RMSE**
Prolonged time of response	75%	0.4	0.89	75%	0.35	0.8
Restricted speech quantity	80%	0.4	0.94	60%	0.6	1
Poor rapport with interviewer	80%	0.4	1	70%	0.45	0.92

Machine (SVM) [160] and Support Vector Regression (SVR) [160], trained in a supervised manner. We used subjective ratings as class labels (1-6) and non-verbal cues as feature-set, then we performed leave-one-person-out cross-validation to calculate the prediction accuracy of each criterion. In Table 5.5 we display only those criteria which could be predicted with more than 60% accuracy. The negative symptoms criteria of *Prolonged Time to Respond*, *Restricted Speech Quantity*, and *Poor Rapport with Interviewer* fall under this category.

To determine whether the subjects and the control cases can be distinguished based on the objective features of non-verbal cues, we utilized binary classifiers in a supervised manner with subjects and controls as training target labels, and performed leave-one-person-out cross-validation to calculate the accuracy of classifica-

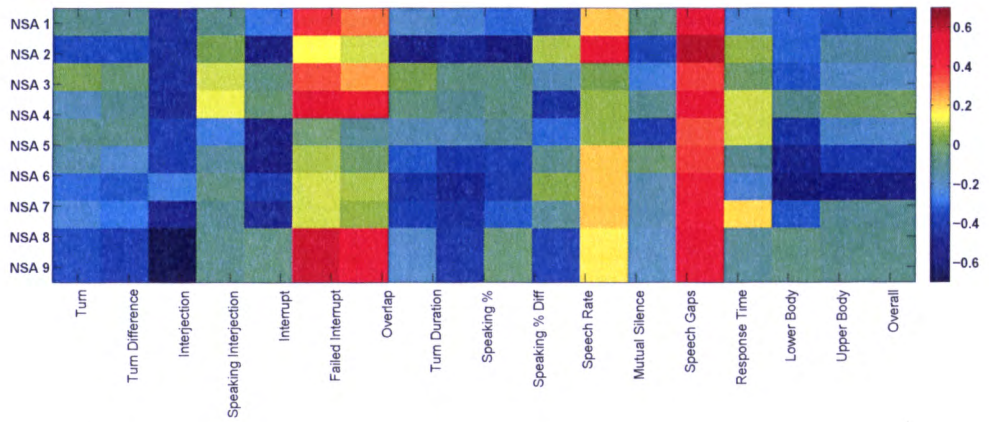


Figure 5.7: Colormap plot between NSA-16 features 1-9 and non-verbal features.

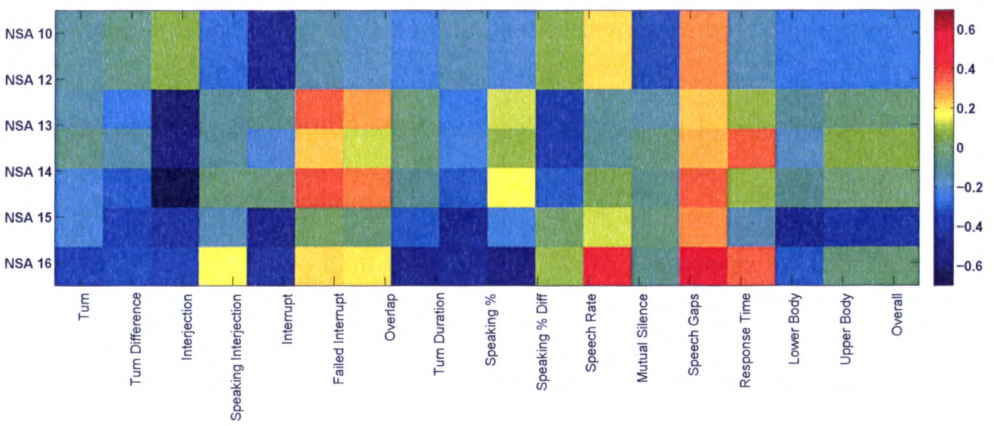


Figure 5.8: Colormap plot between NSA-16 features 10-16 and non-verbal features.

Table 5.6: Classification of conversational speech features into controls and subjects.

Session	Accuracy	MAE	RMSE
Session 1	75%	0.31	0.55
Session 2	79%	0.20	0.45
Session 3	69%	0.25	0.5
Sessions Combined	80%	0.20	0.44

tion. Table 5.6 contains the classification accuracies. We have presented session-wise accuracies as well as combined accuracy for all sessions.

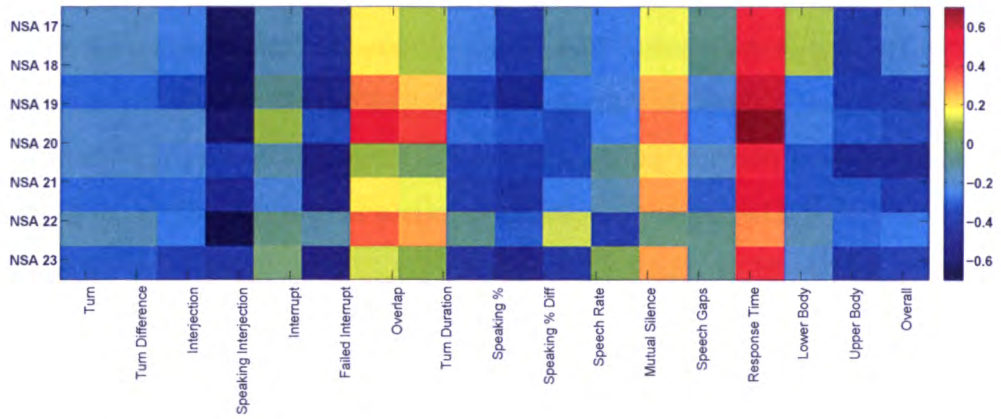


Figure 5.9: Colormap plot between NSA-16 features 17-23 and non-verbal features.

5.4 Discussion

It can be clearly seen from Fig. 5.7, Fig. 5.8 and Fig. 5.9 that strong correlations exist between subjective ratings (NSA-16) and objective measures (Non-verbal cues). An interesting thing to note here is that the correlation values are higher for Figure 5.7 and Figure 5.9 as compared to Figure 5.8. It can be seen from Table 5.4, the NSA-16 questions 1-9 are more related to speech, and hence the absolute correlation values between these subjective questions and the objective speech-related measures are relatively higher compared to those between NSA-16 questions 10-16 and the objective non-verbal speech measures. Similarly, Figure 5.9 has cumulative measures, hence it also has overall higher correlations. The visual features mostly have negative significant correlations with the NSA-16 measures as can be see in Fig. 5.7, Fig. 5.8 and Fig. 5.9.

Response time shows positive correlation with *prolonged response time*, *poor rapport with the interviewer*, *restricted speech quantity*, and *NSA communication*. This relation shows that the response time is higher for the cases where the psychologist rated the patient high for these negative symptoms. *Failed interrupts* shows positive correlation with *poor rapport with the interviewer* and that also makes sense because high interruption usually indicates lack of smoothness in the conversation which can lead to poor rapport with the interviewer. In the case of visual features there are

no significant positive correlations. *Speaking %* and *Interrupts* correlate negatively with *restricted speech quantity.* This relation means that the more a patient speaks, the lower he/she is rated for restricted speech which makes sense. Higher interruption also indicates patients' participation in the conversation therefore less restricted speech quantity. The visual cues mostly correlate negatively with NSA-16 ratings. The most prominent correlation is of *Upper body movement*, and *Overall movement* with *reduced expressive gestures* and this relation makes perfect sense. As the patient is seated so major contributor to the gestures is the upper body, and greater the movement of upper body the lower the rating for *reduced expressive gestures. Upper body movement*, and *Overall movement* also correlate negatively with *affect reduced display on demand* and *emotion affect cumulative rating.*

Few of the strongest correlations are observed for the NSA-16 questions 1,2, and 9, i.e., *Prolonged time to respond*, *Restricted speech quantity*, and *Poor rapport with interviewer* respectively. This is also probably the reason why we achieve high accuracies in predicting the subjective ratings of these questions from our objective speech measures. *Prolonged time to respond*, *Restricted speech quantity*, or *Poor rapport with interviewer* are strongly related to *Response Time* (positive correlation) and to *Interruption* (negative correlation). According to Table 5.4, a patient is rated higher on these questions by the psychologist if the patient has frequent pauses while responding, has responses limited to a few words or has to be constantly prodded to be engaged in the conversation. All these 'anomalies' in the conversation are reflected in the lack of interruptions common in a 'normal' conversation or a increased time to respond to turns in the conversation. These attributes are faithfully captured in our objective speech measures of *Response Time* and *Interruption.* Thus, the subjective perceptions regarding the NSA-16 questions are reaffirmed through objective speech processing techniques.

In Table 5.5, together with the classification accuracies for SVM and SVR techniques, we additionally mention the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) obtained for these classification results. A low value ($\leqslant 1$) on a scale of 1-6 for these measures indicate the error to be quite low.

The results from Table 5.6 clearly indicate that audio features for control and subjects contain major differences and thus could be differentiated with a high accuracy. This difference can be attributed to the difference in schizophrenia severity in the subject and control cases. As mentioned earlier, both the subjects and controls were suffering from schizophrenia, but the subjects were undergoing CRT, whereas the controls were not recommended for CRT by doctors. The controls have relatively superior mental health compared to subjects, and this contrast is also translated into the differences in the non-verbal cues.

Page intentionally left blank

Chapter 6

Conclusion and Future Work

6.1 Conclusion

We presented a novel approach towards comprehensive real-time analysis of sociometrics. We collected two different speech corpora of two-person conversations: an audio corpus and an audio-visual corpus. We leveraged on multi-modal information (audio and video) to determine social behavior in two-person conversations, specifically, interest, dominance, politeness, friendliness, frustration, empathy, respect, confusion, hostility and agreement of the speaker. We asked multiple judges to annotate the recordings in each dataset; these annotations allowed us to train and validate machine learning algorithms for reliable multi-level classification of the sociometrics with speech and visual cues as input features. We performed correlation analysis based on these annotations to determine the positive and negative correlations between various social indicators. This analysis highlighted the redundancies among indicators and provided us with insight regarding the social indicators. We performed separate audio, video and audio-visual analysis and presented the results for each feature set. In order to determine the robustness of the system, we performed a cross dataset analysis. Our results indicate that despite the differences between two data sets, we could still infer sociometrics with reasonable accuracy.

The combined metrics for speech mannerisms and social behavior provide a clear picture of human behavior in dialogs. For instance, high volume, high dominance coupled with low agreement, low politeness and high hostility levels may suggest that the participant is upset and behaving aggressively. Similarly, low volume, low interest and low dominance may suggest that the participant is bored. Low interest from both speakers would mean that both participants are disinterested in the conversation. Similarly, high agreement, high politeness, low hostility, high friendliness and a moderate dominance would mean the conversation is going smoothly. Further research is required to interpret the interplay between these sociometrics.

We designed various feedback platforms for delivering the feedback messages. The aim was to make a sociofeedback system that can be used for real world scenarios. Keeping that in mind we developed VoIP (Skype), Android phones/tablets, Google glass and Vuzix smart glasses based applications that can be interfaced with the sociofeedback system and provide feedback using the inference from the system.

We presented a user study about Sociofeedback via Nao robot. In the first part of the study, we investigated whether users can understand the feedback messages, and compared two modalities for feedback (audio only and audio combined with gestures). To simplify the problem, we used an open-loop setting. The participants did not participate in conversations; instead they only needed to assess the feedback messages delivered by the NAO robot. Each participant completed a (modified) Godspeed questionnaire twice, once to assess audio-only messages, and once where audio and gestures are combined. We observed that the feedback was identified more accurately when audio and gestures are combined, leading to (almost) perfect identification. The results of the (modified) Godspeed questionnaire suggest that by combing audio with gestures in sociofeedback, the anthropomorphism, animacy, likeability and perceived intelligence of the NAO robot increases, as compared to feedback through audio messages only. The ratings for all Godspeed criteria were high. In conclusion, this study suggests that sociofeedback by the NAO robot can be accurately identified, and is also appreciated by participants.

In the second part of the study, we delivered feedback for dyadic conversations via Nao. In this experiment, we used a close-loop setting. The participants took part in real conversation; the Nao robot monitored these conversations, and provided feedback to the participants regarding their social behavior. We aimed to investigate how the feedback from the humanoid robot is perceived by humans. To this end, we conducted a survey with 20 participants where the participants were engaged in a discussion, and the feedback messages were delivered by Nao to the participants. The participants assessed the content, timing, relevance and their liking of the feedback after receiving each feedback message.

We observed that the participants clearly liked receiving feedback from Nao robot. The agreement scores are very high, showing that the participants agreed with the provided feedback. There is room for improvement in the timing of the feedback. We will try to improve the timing in future experiments.

At the end, each participant assessed the robot and rated it on a Godspeed questionnaire. The ratings for all Godspeed criteria were high implying that participants liked receiving feedback from a humanoid robot. Only with regard to perceived safety, the evaluation was mildly positive; this may be explained by the fact that the robot was sitting near the participants. However, the average rating is still acceptable, and this issue may not be very critical. Overall, this study suggests that sociofeedback by the Nao robot can be accurately identified and is appreciated by participants.

We also presented our findings regarding the correlations between the non-verbal speech and visual cues and negative symptom ratings. Positive symptoms of schizophrenia can be treated using medication. However, there is no treatment for the negative symptoms. At Institute for Mental Health, Singapore doctors administered cognitive remediation therapy (CRT) to schizophrenia patients. These patients were interviewed and rated on negative symptoms over the course of their treatment. We recorded these interviews and extracted non-verbal audio features.

The results of our analysis are promising as there are significant correlations be-

tween non-verbal features and NSA-16 ratings assigned by psychologist. We also predicted NSA-16 criteria using machine learning algorithms trained on subjective ratings. The results show that some of the NSA-16 criteria can be predicted using non-verbal features with quite high accuracy. The aim of this work is to exploit these correlations and predict the negative symptoms criteria in an objective manner. These results can be the stepping stone towards building an automated tool which could predict negative symptoms by analyzing the speech of a patient in an automated manner. Such tool may serve as an aid to psychologists, and could potentially help them in providing better monitoring of schizophrenia patients. It can be extremely helpful in the area of telemedicine for patients suffering from schizophrenia, where treatment can be administered remotely and objective metrics for the patients' speech behavior can assist the psychologists in accurately providing the ratings. These are promising, albeit based only on data from a relatively small number of patients.

In summary we demonstrated that sociometrics can be computed in a quick and reliable manner, enabling real-time feedback ("sociofeedback"), thereby closing the research gap in applying sociometric results to build a real time system. In addition, we presented a real world implementation where sociofeedback system was interfaced with Nao robot enabling Nao robot to deliver feedback for dyad conversations. Currently, we are exploring various applications of sociofeedback. We also presented a rehabilitative application for schizophrenic patients. Our results show that a correlation does exist between non-verbal cues and negative schizophrenia symptoms.

6.2 Future Work

6.2.1 Sociometrics

Our ongoing work includes the extension of two-person dialogs to multi-party dialogs. We intend to collect a larger and diverse dataset in order to generalize the

findings. The number of speakers does not pose a big challenge as we use separate audio channels for each speaker. We use one Kinect per person in our current setting. However, Kinect SDK can detect multiple skeletons using one device. Hence, we can technically reduce the numbers of Kinect devices for multiparty scenarios.

We also need to adapt our audio-visual cues for multiparty settings. The dynamics of the multiparty conversations is very different from that of dyadic conversation. The sociofeedback system generates feedback for each speaker and hence, the analysis has to be done accordingly. For dyadic conversations, detecting speaking/non-speaking parts from the speech data was sufficient to determine the non-verbal cues. In the case of multiparty discussions, it is necessary to determine whom the speaker is addressing in order to extract conversational cues to generate suitable feedback.

We also plan to incorporate context detection in the sociofeedback system. We believe context detection can enhance the robustness of the system. To achieve this, we are working towards integrating real-time speech and object detection. Speech detection can provide us with keywords that can help us understand the context. Similarly, object detection can help with the understanding of the background surroundings, clothes, gender, ethnicity etc. All this additional information can help us improve the sociofeedback system.

6.2.2 Feedback Platforms

To date, we have only implemented the applications to present the feedback messages on various platforms such as VoIP (Skype), Android phone/tablet and smartglasses. Currently, we implement a TCP server/client framework where the sociofeedback system sends the inferred social state to the application. The application then displays the feedback in an appropriate format.

In future work, we plan on conducting user studies similar to the one in chapter 4 to get user opinion about these feedback platforms. We also plan to implement the entire Sociofeedback system on Android platform, in which case we will not need a

TCP server to provide feedback messages. The Android device will then store and process the audio, and infer the social state of the speaker using trained classifiers. Android platforms already have the technical specifications. We are working on an app that can independently acquire speech data, extract speech cues, infer the social state and provide feedback. This will open new avenues for research. Our current focus has been on dyad conversations. The independent mobile platform can help us apply similar techniques to an individual speaker. Wearable devices such as smart glasses can be used for training individuals to improve presentation skills, sales pitches, teaching etc.

There are technical challenges involved with the implementation of the sociofeedback system on VoIP and Android platforms. Currently, we use separate audio channel for each speaker, which is not feasible on a mobile platform. We plan to use speaker diarization to separate speakers audio from a conversation. Once we have separated the audio, we can extract speaking/non-speaking sections and then extract non-verbal cues. We use Microsoft Kinect sensor to collect visual and depth data. In VoIP (Skype), we can still use Kinect which can serve as a web cam, making it possible to acquire depth data while the user is in a Skype call. For the Android platform video data, it will be a challenge, as acquiring depth data is not possible with existing technology. However, if the user keeps the phone/tablet in front of his/her face, we can extract facial expressions and context related features.

6.2.3 Social Robotics

The ability for a robot to estimate the social state of the speaker with whom it is interacting can enhance human robot interaction. The user studies presented in our work cover the scenario where a humanoid robot provides sociofeedback for dyadic conversations.

In future, we plan to develop a generic social state estimation module for humanoid robots. The action of the robot to the estimated social state would vary based on the application. In the current setting, the robot only provides feedback. But, in case

it is a party to the conversation then it can generate dialogs based on the estimated social state. It would be ideal to have a system that can infer the social state and provide that data to the developers. The developers can then program the action based on their application requirements.

6.2.4 Non-verbal Analysis of Schizophrenic patients' Interviews

This study is currently ongoing. Hence, we have only presented the initial results in this thesis. The work presented here focuses on the correlations between the non-verbal speech and visual cues with negative schizophrenia symptoms rated by a trained psychologist. In the future, we plan to increase the number of participants (both subjects and controls) to obtain more reliable results.

Additionally, we wish to explore the variation of the non-verbal cues for the subjects and controls, specifically, how their speech features change over sessions. If these cues change differently for subjects and controls, this can shed light over the efficacy of Cognitive Remediation Therapy (CRT).

We have also started data collection for healthy individuals. Once we have substantial data, we can use it to compare the non-verbal cues of schizophrenic patients with healthy people. This comparison can give us valuable insight in understanding the speech patterns of schizophrenic patients in comparison to healthy people.

Currently, there is no system that can predict the subjective NSA-16 ratings used for schizophrenia treatment. Our results show that some of these measures can be predicted with reasonable accuracies using non-verbal cues. We plan to exploit these relations and develop an application that can automatically judge NSA-16 criteria and help the schizophrenic patients with their rehabilitation.

Page intentionally left blank

Publications

1. Tahir, Y., Chakraborty, D., Maszczyk, T., Dauwels, S., Dauwels, J., Thalmann, N. and Thalmann, D., 2015, July. Real-time sociometrics from audio-visual features for two-person dialogs. In 2015 IEEE International Conference on Digital Signal Processing (DSP) (pp. 823-827). IEEE.

2. Tahir, Y., Rasheed, U., Dauwels, S. and Dauwels, J., 2014, March. Perception of humanoid social mediator in two-person dialogs. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (pp. 300-301). ACM.

3. Tahir, Y., Rasheed, U., Hui, K., Dauwels, S., Dauwels, J., Thalmann, D. and Thalmann, N.M., 2013, October. NAO Robot as a Social Mediator: A User Study. In International Conference on Social Robotics (ICSR2013), Bristol, UK.

4. Tahir, Y., Chakraborty, D., Dauwels, J., Thalmann, N., Thalmann, D. and Lee, J., 2016, March. Non-verbal speech analysis of interviews with schizophrenic patients. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5810-5814). IEEE.

5. Rasheed, U., Tahir, Y., Dauwels, S., Dauwels, J., Thalmann, D. and Magnenat-Thalmann, N., 2013, October. Real-time comprehensive sociometrics for two-person dialogs. In International Workshop on Human Behavior Understanding (pp. 196-208). Springer International Publishing.

6. Sarda, S., Constable, M., Dauwels, J., Dauwels, S., Elgendi, M., Mengyu, Z., Rasheed, U., Tahir, Y., Thalmann, D. and Magnenat-Thalmann, N., 2014. Real-time feedback system for monitoring and facilitating discussions. In Natural Interaction with Robots, Knowbots and Smartphones (pp. 375-387). Springer New York.

7. Constable, M., Dauwels, J., Dauwels, S., Umer, R., Zhou, M., Tahir, Y. (2016). Modelling Conversation. In Context Aware Human-Robot and Human-Agent Interaction (pp. 81-111). Springer International Publishing.

Bibliography

[1] R. R. Hassin, J. S. Uleman, and J. A. Bargh, *The new unconscious.* Oxford University Press, 2005, vol. 1.

[2] M. Knapp, J. Hall, and T. Horgan, *Nonverbal communication in human interaction.* Cengage Learning, 2013.

[3] V. Manusov and M. L. Patterson, *The Sage handbook of nonverbal communication.* Sage Publications, 2006.

[4] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69–87, 2012.

[5] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[6] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," in *AAAI/IAAI*, 2000, pp. 679–684.

[7] Y. Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* Association for Computational Linguistics, 2006, pp. 81–84.

[8] B. Hutchinson, B. Zhang, and M. Ostendorf, "Unsupervised broadcast conversation speaker role labeling," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* IEEE, 2010, pp. 5322–5325.

[9] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, "From nonverbal cues to perception: personality and social attractiveness," in *Cognitive Behavioural Systems.* Springer, 2012, pp. 60–72.

[10] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri, "Multimodal support to group dynamics," *Personal and Ubiquitous Computing*, vol. 12, no. 3, pp. 181–195, 2008.

[11] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: mapping nonverbal vocal behavior into trait attributions," in *Proceedings of the 2nd international workshop on Social signal processing.* ACM, 2010, pp. 17–20.

[12] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013.

[13] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge." in *INTERSPEECH*, 2012.

[14] R. Nishimura, N. Kitaoka, and S. Nakagawa, "Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling." in *INTERSPEECH*, 2008, pp. 534–537.

[15] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction.* Springer, 2005, pp. 181–200.

[16] B. Hornler and G. Rigoll, "Multi-modal activity and dominance detection in smart meeting rooms," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.* IEEE, 2009, pp. 1777–1780.

[17] L. S. Kennedy and D. P. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on.* IEEE, 2003, pp. 243–248.

[18] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP).* Citeseer, 2005, pp. 489–492.

[19] W. Wang, S. Yaman, K. Precoda, and C. Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 5556–5559.

[20] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: ratings and analysis of broadcast political debates," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE, 2012, pp. 5089–5092.

[21] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2.* Association for Computational Linguistics, 2003, pp. 34–36.

[22] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools," *Image and Vision Computing*, vol. 31, no. 2, pp. 203–221, 2013.

[23] K. Bousmalis, L.-P. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on.* IEEE, 2011, pp. 746–752.

[24] A. Vinciarelli, "Capturing order in social interactions [social sciences]," *IEEE Signal Processing Magazine*, vol. 26, no. 5, pp. 133–152, 2009.

[25] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 2004, p. 669.

[26] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *Proceedings of 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, 2003 (ICASSP'03)*, vol. 1. IEEE, 2003, pp. I–364.

[27] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[28] O. Aran, H. Hung, and D. Gatica-Perez, "A multimodal corpus for studying dominance in small group conversations," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality 18 May 2010*, p. 22, 2010.

[29] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, "Detection and application of influence rankings in small group meetings," in *Proceedings of the 8th international conference on Multimodal interfaces.* ACM, 2006, pp. 257–264.

[30] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 501–513, 2009.

[31] N. E. Dunbar and J. K. Burgoon, "Perceptions of power and interactional dominance in interpersonal relationships," *Journal of Social and Personal Relationships*, vol. 22, no. 2, pp. 207–233, 2005.

[32] M. S. Mast, "Dominance as expressed and inferred through speaking time," *Human Communication Research*, vol. 28, no. 3, pp. 420–450, 2002.

[33] K. J. Tusing and J. P. Dillard, "The sounds of dominance." *Human Communication Research*, vol. 26, no. 1, pp. 148–171, 2000.

[34] O. Aran and D. Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in group conversations," in *Pattern Recognition (ICPR), 2010 20th International Conference on.* IEEE, 2010, pp. 3687–3690.

[35] K. Kalimeri, B. Lepri, O. Aran, D. B. Jayagopi, D. Gatica-Perez, and F. Pianesi, "Modeling dominance effects on nonverbal behaviors using granger causality," in *Proceedings of the 14th ACM international conference on Multimodal interaction.* ACM, 2012, pp. 23–26.

[36] W. Wang, K. Precoda, R. Hadsell, Z. Kira, C. Richey, and G. Jiva, "Detecting leadership and cohesion in spoken interactions," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE, 2012, pp. 5105–5108.

[37] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *Multimedia, IEEE Transactions on*, vol. 14, no. 3, pp. 816–832, 2012.

[38] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, and D. Gatica-Perez, "Body communicative cue extraction for conversational analysis," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on.* IEEE, 2013, pp. 1–8.

[39] M. Mahmoud, T. Baltrušaitis, P. Robinson, and L. D. Riek, "3d corpus of spontaneous complex mental states," in *Affective Computing and Intelligent Interaction.* Springer, 2011, pp. 205–214.

[40] G. Galatas, G. Potamianos, D. Kosmopoulos, C. McMurrough, and F. Makedon, "Bilingual corpus for avasr using multiple sensors and depth information," in *Proc. AVSP*, 2011, pp. 103–106.

[41] U. Rasheed, Y. Tahir, S. Dauwels, J. Dauwels, D. Thalmann, and N. Magnenat-Thalmann, "Real-time comprehensive sociometrics for two-person dialogs," in *Human Behavior Understanding.* Springer, 2013, pp. 196–208.

[42] Y. Tahir, D. Chakraborty, T. Maszczyk, S. Dauwels, J. Dauwels, N. Thalmann, and D. Thalmann, "Real-time sociometrics from audio-visual features for two-person dialogs," in *Digital Signal Processing (DSP), 2015 IEEE International Conference on.* IEEE, 2015, pp. 823–827.

[43] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction.* Springer, 2006, pp. 28–39.

[44] N. Eagle and A. S. Pentland, "Social network computing," in *UbiComp 2003: Ubiquitous Computing.* Springer, 2003, pp. 289–296.

[45] J. Sturm, O. H.-v. Herwijnen, A. Eyck, and J. Terken, "Influencing social dynamics in meetings through a peripheral display," in *Proceedings of the 9th international conference on Multimodal interfaces.* ACM, 2007, pp. 263–270.

[46] T. Kim, A. Chang, L. Holland, and A. S. Pentland, "Meeting mediator: enhancing group collaborationusing sociometric feedback," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 2008, pp. 457–466.

[47] J. M. DiMicco, A. Pandolfo, and W. Bender, "Influencing group participation with a shared display," in *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, 2004, pp. 614–623.

[48] K. Bachour, F. Kaplan, and P. Dillenbourg, "Reflect: An interactive table for regulating face-to-face collaborative learning," in *Times of Convergence. Technologies Across Learning Contexts*. Springer, 2008, pp. 39–48.

[49] A. Pentland and S. Pentland, "Honest signals: how they shape our world. 2008," *Boston: Massachusetts Institute of Technology*.

[50] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 143–166, 2003.

[51] H. Li, J.-J. Cabibihan, and Y. K. Tan, "Towards an effective design of social robots," *International Journal of Social Robotics*, vol. 3, no. 4, pp. 333–335, 2011.

[52] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "A communication robot in a shopping mall," *IEEE Transactions on Robotics*, vol. 26, no. 5, pp. 897–913, 2010.

[53] S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz, "Minerva: A second-generation museum tour-guide robot," in *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA 99)*, 1999.

[54] F. Tanaka, A. Cicourel, and J. R. Movellan, "Socialization between toddlers and robots at an early childhood education center," *Proceedings of the National Academy of Science*, vol. 104, pp. 17 954–17 958, Nov. 2007.

[55] B. Graf, C. Parlitz, and M. Hägele, "Robotic home assistant care-o-bot® 3 product vision and innovation platform," in *HCI (2)*, 2009, pp. 312–320.

[56] K. Williams and C. Breazeal, "A reasoning architecture for human-robot joint tasks using physics-, social-, and capability-based logic," in *IROS*, 2012, pp. 664–671.

[57] D. François, D. Polani, and K. Dautenhahn, "Towards socially adaptive robots: A novel method for real time recognition of human-robot interaction styles," in *Humanoids*, 2008, pp. 353–359.

[58] F. Papadopoulos, K. Dautenhahn, and W. C. Ho, "Exploring the use of robots as social mediators in a remote human-human collaborative communication experiment," *Paladyn*, vol. 3, no. 1, pp. 1–10, 2012.

[59] K. Dautenhahn, "Socially intelligent robots: dimensions of human–robot interaction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 679–704, 2007.

[60] D. Bohus and E. Horvitz, "Dialog in the open world: platform and applications," in *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 2009, pp. 31–38.

[61] A. A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, *Human Behavior Understanding: First International Workshop, HBU 2010, Istanbul, Turkey, August 22, 2010, Proceedings*. Springer, 2010, vol. 6219.

[62] A. A. Salah, J. Ruiz-del Solar, Ç. Meriçli, and P.-Y. Oudeyer, "Human behavior understanding for robotics," in *Human Behavior Understanding*. Springer, 2012, pp. 1–16.

[63] A. S. Pentland, *Honest signals*. MIT press, 2010.

[64] Y. Tahir, U. Rasheed, S. Dauwels, J. Dauwels, N. Thalmann, and D. Thalmann, "Nao as social mediator: A user study," in *Robots in public*

spaces:towards multi-party, short-term, dynamic human-robot interaction. Springer, 2013.

[65] Y. Tahir, U. Rasheed, S. Dauwels, and J. Dauwels, "Perception of humanoid social mediator in two-person dialogs," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction.* ACM, 2014, pp. 300–301.

[66] C. Bartneck, E. Croft, and D. Kulic, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots," in *Metrics for HRI Workshop, Technical Report*, vol. 471. Citeseer, 2008, pp. 37–44.

[67] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *rn*, vol. 255, p. 3, 1999.

[68] C. L. Sidner and M. Dzikovska, "A first experiment in engagement for human-robot interaction in hosting activities," in *Advances in Natural Multimodal Dialogue Systems.* Springer, 2005, pp. 55–76.

[69] M. E. Pollack, L. Brown, D. Colbry, C. Orosz, B. Peintner, S. Ramakrishnan, S. Engberg, J. T. Matthews, J. Dunbar-Jacob, C. E. McCarthy *et al.*, "Pearl: A mobile robotic assistant for the elderly," in *AAAI workshop on automation as eldercare*, vol. 2002, 2002, pp. 85–91.

[70] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu, "Robovie: an interactive humanoid robot," *Industrial robot: An international journal*, vol. 28, no. 6, pp. 498–504, 2001.

[71] A. Billard, "Robota: Clever toy and educational tool," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 259–269, 2003.

[72] C. D. Kidd, W. Taggart, and S. Turkle, "A sociable robot to encourage social interaction among the elderly," in *Robotics and Automation, 2006. ICRA*

2006. Proceedings 2006 IEEE International Conference on. IEEE, 2006, pp. 3972–3976.

[73] K. C. Welch, U. Lahiri, Z. Warren, and N. Sarkar, "An approach to the design of socially acceptable robots for children with autism spectrum disorders," *International Journal of Social Robotics*, vol. 2, no. 4, pp. 391–403, 2010.

[74] I. Fujimoto, T. Matsumoto, P. R. S. De Silva, M. Kobayashi, and M. Higashi, "Mimicking and evaluating human motion to improve the imitation skill of children with autism through a robot," *International Journal of Social Robotics*, vol. 3, no. 4, pp. 349–357, 2011.

[75] G. Schiavone, D. Formica, F. Taffoni, D. Campolo, E. Guglielmelli, and F. Keller, "Multimodal ecological technology: From childs social behavior assessment to child-robot interaction improvement," *International Journal of Social Robotics*, vol. 3, no. 1, pp. 69–81, 2011.

[76] J.-J. Cabibihan, H. Javed, M. Ang Jr, and S. M. Aljunied, "Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism," *International Journal of Social Robotics*, pp. 1–26, 2013.

[77] A. Powers and S. Kiesler, "The advisor robot: tracing people's mental model from a robot's physical attributes," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction.* ACM, 2006, pp. 218–225.

[78] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. Petrick, "Two people walk into a bar: Dynamic multi-party social interaction with a robot agent," in *Proceedings of the 14th ACM international conference on Multimodal interaction.* ACM, 2012, pp. 3–10.

[79] G. Tur, A. Stolcke, L. Voss, S. Peters, D. Hakkani-Tur, J. Dowding, B. Favre, R. Fernández, M. Frampton, M. Frandsen *et al.*, "The calo meeting assistant

system," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1601–1611, 2010.

[80] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive Behavioural Systems*. Springer, 2012, pp. 114–130.

[81] J. Harris and E. Sharlin, "Exploring the affect of abstract motion in social human-robot interaction," in *RO-MAN, 2011 IEEE*. IEEE, 2011, pp. 441–448.

[82] M. Saerbeck and C. Bartneck, "Perception of affect elicited by robot motion," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2010, pp. 53–60.

[83] A. Beck, A. Hiolle, A. Mazel, and L. Cañamero, "Interpretation of emotional body language displayed by robots," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*. ACM, 2010, pp. 37–42.

[84] J. Ham, R. Bokhorst, and J. Cabibihan, "The influence of gazing and gestures of a storytelling robot on its persuasive power," in *International conference on social robotics*, 2011.

[85] A. Delaborde and L. Devillers, "Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*. ACM, 2010, pp. 75–80.

[86] N. C. Andreasen and S. Olsen, "Negative v positive schizophrenia: definition and validation," *Archives of General Psychiatry*, vol. 39, no. 7, pp. 789–794, 1982.

[87] C. Demily and N. Franck, "Cognitive remediation: a promising tool for the treatment of schizophrenia," 2008.

[88] H. Brenner, S. Kraemer, M. Hermanutz, and B. Hodel, "Cognitive treatment in schizophrenia," in *Schizophrenia*. Springer, 1990, pp. 161–191.

[89] A. G. Cardno, E. J. Marshall, B. Coid, A. M. Macdonald, T. R. Ribchester, N. J. Davies, P. Venturi, L. A. Jones, S. W. Lewis, P. C. Sham *et al.*, "Heritability estimates for psychotic disorders: the maudsley twin psychosis series," *Archives of general psychiatry*, vol. 56, no. 2, pp. 162–168, 1999.

[90] U. Heresco-Levy, D. C. Javitt, M. Ermilov, C. Mordel, G. Silipo, and M. Lichtenstein, "Efficacy of high-dose glycine in the treatment of enduring negative symptoms of schizophrenia," *Archives of general psychiatry*, vol. 56, no. 1, pp. 29–36, 1999.

[91] U. Heresco-Levy, M. Ermilov, P. Lichtenberg, G. Bar, and D. C. Javitt, "High-dose glycine added to olanzapine and risperidone for the treatment of schizophrenia," *Biological psychiatry*, vol. 55, no. 2, pp. 165–171, 2004.

[92] D. C. Goff, G. Tsai, D. S. Manoach, and J. T. Coyle, "Dose-finding trial of d-cycloserine added to neuroleptics for negative symptoms in schizophrenia," *The American journal of psychiatry*, vol. 152, no. 8, p. 1213, 1995.

[93] D. C. Goff, G. Tsai, J. Levitt, E. Amico, D. Manoach, D. A. Schoenfeld, D. L. Hayden, R. McCarley, and J. T. Coyle, "A placebo-controlled trial of d-cycloserine added to conventional neuroleptics in patients with schizophrenia," *Archives of General Psychiatry*, vol. 56, no. 1, pp. 21–27, 1999.

[94] R. W. Buchanan, D. C. Javitt, S. R. Marder, N. R. Schooler, J. M. Gold, R. P. McMahon, M. Uriel Heresco-Levy, and W. T. Carpenter, "The cognitive and negative symptoms in schizophrenia trial (consist): the efficacy of glutamatergic agents for negative symptoms and cognitive impairments," *The American journal of psychiatry*, vol. 164, no. 10, pp. 1593–1602, 2007.

[95] B. P. Murphy, Y.-C. Chung, T.-W. Park, and P. D. McGorry, "Pharmacological treatment of primary negative symptoms in schizophrenia: a systematic review," *Schizophrenia research*, vol. 88, no. 1, pp. 5–25, 2006.

[96] C. A. Tamminga, R. W. Buchanan, and J. M. Gold, "The role of negative symptoms and cognitive dysfunction in schizophrenia outcome." *International clinical psychopharmacology*, 1998.

[97] P. Milev, B.-C. Ho, S. Arndt, and N. C. Andreasen, "Predictive values of neurocognition and negative symptoms on functional outcome in schizophrenia: a longitudinal first-episode study with 7-year follow-up," *American Journal of Psychiatry*, vol. 162, no. 3, pp. 495–506, 2005.

[98] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[99] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009. Citeseer, 2009, pp. 312–315.

[100] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge." in *INTERSPEECH*, vol. 2010, 2010, pp. 2795–2798.

[101] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge." in *INTERSPEECH*, 2011, pp. 3201–3204.

[102] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge." in *INTERSPEECH*, vol. 2012, 2012, pp. 254–257.

[103] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech

2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.

[104] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The interspeech 2014 computational paralinguistics challenge: cognitive & physical load." in *INTERSPEECH*, 2014, pp. 427–431.

[105] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinsons & eating condition," in *Proceedings of Interspeech*, 2015.

[106] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge.* ACM, 2013, pp. 3–10.

[107] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.* ACM, 2014, pp. 3–10.

[108] L. E. DeLisi, "Speech disorder in schizophrenia: Review of the literature and exploration of its relation to the uniquely human capacity for language." *Schizophrenia Bulletin*, vol. 27, no. 3, p. 481, 2001.

[109] B. Elvevåg, D. Weinstock, M. Akil, J. Kleinman, and T. Goldberg, "A comparison of verbal fluency tasks in schizophrenic patients and normal controls," *Schizophrenia Research*, vol. 51, no. 2, pp. 119–126, 2001.

[110] B. Elvevåg, T. Weickert, M. Wechsler, R. Coppola, D. Weinberger, and T. Goldberg, "An investigation of the integrity of semantic boundaries in schizophrenia," *Schizophrenia Research*, vol. 53, no. 3, pp. 187–198, 2002.

[111] B. Elvevåg, K. Helsen, M. De Hert, K. Sweers, and G. Storms, "Metaphor interpretation and use: a window into semantics in schizophrenia," *Schizophrenia research*, vol. 133, no. 1, pp. 205–211, 2011.

[112] S. McGilloway, S. J. Cooper, and E. Douglas-Cowie, "Can patients with chronic schizophrenia express emotion? a speech analysis," *Schizophrenia research*, vol. 64, no. 2, pp. 189–190, 2003.

[113] B. Elvevåg, P. W. Foltz, M. Rosenstein, and L. E. DeLisi, "An automated method to analyze language use in patients with schizophrenia and their first-degree relatives," *Journal of neurolinguistics*, vol. 23, no. 3, pp. 270–284, 2010.

[114] K. Holshausen, P. D. Harvey, B. Elvevåg, P. W. Foltz, and C. R. Bowie, "Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia," *Cortex*, vol. 55, pp. 88–96, 2014.

[115] G. Bedi, F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, and C. M. Corcoran, "Automated analysis of free speech predicts psychosis onset in high-risk youths," *npj Schizophrenia*, vol. 1, 2015.

[116] A. Troisi, G. Spalletta, and A. Pasini, "Non-verbal behaviour deficits in schizophrenia: an ethological study of drug-free patients," *Acta Psychiatrica Scandinavica*, vol. 97, no. 2, pp. 109–115, 1998.

[117] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech." in *Interspeech*, 2013, pp. 857–861.

[118] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd." in *Interspeech*, 2013, pp. 847–851.

[119] D. E. Sturim, P. A. Torres-Carrasquillo, T. F. Quatieri, N. Malyska, and A. McCree, "Automatic detection of depression in speech using gaussian mixture modeling with factor analysis." in *Interspeech*, 2011, pp. 2981–2984.

[120] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues." in *Interspeech*, 2013, pp. 1149–1153.

[121] G. An, D. G. Brizan, M. Ma, M. Morales, A. R. Syed, and A. Rosenberg, "Automatic recognition of unified parkinsons disease rating from speech with acoustic, i-vector and phonotactic features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[122] J. Kim, M. Nasir, R. Gupta, M. V. Segbroeck, D. Bone, M. Black, Z. I. Skordilis, Z. Yang, P. Georgiou, and S. Narayanan, "Automatic estimation of parkinsons disease severity from diverse speech tasks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[123] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist." in *INTERSPEECH*, 2012, pp. 1043–1046.

[124] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders." in *INTERSPEECH*, 2013, pp. 191–194.

[125] T. Chaspari, C.-C. Lee, and S. Narayanan, "Interplay between verbal response latency and physiology of children with autism during eca interactions." in *INTERSPEECH*, 2012, pp. 1319–1322.

[126] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response

collected via interactive voice response (ivr) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[127] A. F. Leentjens, F. R. Verhey, R. Lousberg, H. Spitsbergen, and F. W. Wilmink, "The validity of the hamilton and montgomery-åsberg depression rating scales as screening and diagnostic tools for depression in parkinson's disease," *International journal of geriatric psychiatry*, vol. 15, no. 7, pp. 644–649, 2000.

[128] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber *et al.*, "The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression," *Biological psychiatry*, vol. 54, no. 5, pp. 573–583, 2003.

[129] N. C. Andreasen, "Negative symptoms in schizophrenia: definition and reliability," *Archives of General Psychiatry*, vol. 39, no. 7, pp. 784–788, 1982.

[130] M. Lavelle, S. Dimic, C. Wildgrube, R. McCabe, and S. Priebe, "Non-verbal communication in meetings of psychiatrists and patients with schizophrenia," *Acta Psychiatrica Scandinavica*, vol. 131, no. 3, pp. 197–205, 2015.

[131] S. Basu, "A linked-hmm model for robust voicing and speech detection," in *Proceedings of 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, 2003 (ICASSP'03)*, vol. 1. IEEE, 2003, pp. I–816.

[132] P. Briñol, R. E. Petty, and B. Wagner, "Body posture effects on self-evaluation: A self-validation approach," *European Journal of Social Psychology*, vol. 39, no. 6, pp. 1053–1064, 2009.

[133] (2010) Kinect sdk windows. [Online]. Available: https://www.microsoft.com/en-us/kinectforwindows/

[134] O. Komarov, "Schemaball," [Software], June 2013, available from http://www.mathworks.com/matlabcentral/fileexchange/42279-schemaball.

[135] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing).* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[136] K. Deng, "Omega: On-line memory-based general purpose system classifier," Ph.D. dissertation, Georgia Institute of Technology, 1998.

[137] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning.* MIT press, 2012.

[138] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural computation*, vol. 19, no. 3, pp. 792–815, 2007.

[139] Mathworks. [Online]. Available: http://www.mathworks.com/

[140] (2003) Spider machine learning toolbox. [Online]. Available: http://people.kyb.tuebingen.mpg.de/spider/

[141] Skype. [Online]. Available: http://www.skype.com/en/

[142] Gotomeeting. [Online]. Available: http://www.gotomeeting.com.sg/

[143] Viber. [Online]. Available: http://www.viber.com/en/

[144] (2015) Sociofeedback demo. [Online]. Available: https://www.youtube.com/watch?v=3uT9O3MqUOg

[145] Supertintin best skype video call recorder. [Online]. Available: http://www.supertintin.com/

[146] Skype4com event handler. [Online]. Available: http://www.codeproject.com/Articles/31009/Skype-COM-Event-Handler-Example-for-ALL-Sky

[147] Introduction to android. [Online]. Available: https://developer.android.com/guide/index.html

[148] (2013) Vuzix m100 smartglasses. [Online]. Available: http://www.vuzix.com/consumer/products_m100/

[149] (2013) Google glass. [Online]. Available: https://developers.google.com/glass/

[150] K. W. Brown and R. M. Ryan, "The benefits of being present: mindfulness and its role in psychological well-being." *Journal of personality and social psychology*, vol. 84, no. 4, p. 822, 2003.

[151] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: A lazy learning approach," 2000.

[152] "Neural networks-a comprehensive foundation neural networks-a comprehensive foundation, 1994."

[153] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM New York, 2001, pp. 41–46.

[154] K. Werner, J. Oberzaucher, and F. Werner, "Evaluation of human robot interaction factors of a socially assistive robot together with older people," in *Complex, Intelligent and Software Intensive Systems (CISIS), 2012 Sixth International Conference on.* IEEE, 2012, pp. 455–460.

[155] S. R. McGurk, E. W. Twamley, D. I. Sitzer, G. J. McHugo, and K. T. Mueser, "A meta-analysis of cognitive remediation in schizophrenia," *The American journal of psychiatry*, vol. 164, no. 12, pp. 1791–1802, 2007.

[156] R. S. Keefe, T. E. Goldberg, P. D. Harvey, J. M. Gold, M. P. Poe, and L. Coughenour, "The brief assessment of cognition in schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery," *Schizophrenia research*, vol. 68, no. 2, pp. 283–297, 2004.

[157] B. N. Axelrod, R. S. Goldman, and L. D. Alphs, "Validation of the 16-item negative symptom assessment," *Journal of psychiatric research*, vol. 27, no. 3, pp. 253–258, 1993.

[158] K. D. Cicerone, D. M. Langenbahn, C. Braden, J. F. Malec, K. Kalmar, M. Fraas, T. Felicetti, L. Laatsch, J. P. Harley, T. Bergquist *et al.*, "Evidence-based cognitive rehabilitation: updated review of the literature from 2003 through 2008," *Archives of physical medicine and rehabilitation*, vol. 92, no. 4, pp. 519–530, 2011.

[159] T. Wykes, V. Huddy, C. Cellard, S. R. McGurk, and P. Czobor, "A meta-analysis of cognitive remediation for schizophrenia: methodology and effect sizes," *American Journal of Psychiatry*, 2014.

[160] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

.1 Transcripts of conversations

Dialog 1:- (An interaction between a salesperson and a dissatisfied customer.)

A:- Morning sir how may I help you?

B:- I am here to return this phone.

A:- Sir, we cannot return sold items.

B:- What do you mean you cannot return? You sold me a broken phone and now you cannot return!

A:- Sir please listen to me.

B:- Just call the manager, I will talk to him.

A:- Sir, please just tell me what is the problem.

B:- There is a complete list. I am shocked to see the difference between your claims and this phones performance! The battery does not charge properly, the screen flickers, headphones dont work. Do you want me to go on?

A:- Sir, kindly tell me clearly?

B:- Just call the manager.

A:- Sir, the manager is not available right now.

B:- What do you mean he is not available? I am here in working hours, he should be here. What kind of an irresponsible business are you?

A:- He is attending to an emergency. I will try to call him.

B:- So, your clients problems do not mean anything! I have wasted my time and money on this phone and you do not have the decency to attend to my complaints!

A:- Kindly leave your contact details. We will contact you.

B:- No need for that, just keep this broken phone and I will buy from someone who knows how to run a business.

A:- Sir please understand.

B:- I am understanding. You guys should learn how to run a business. Improve your service or your business will close. This is no way to run a business, I am telling you.

A:- There is no need to be angry sir.

B:- Well I am the one who has a broken phone. Have you seen your competitors in the market? They are doing much better because their service is good. Learn something from them.

Dialog 2:- (A small dialog between two uninterested speakers)

A:- Good Morning.

B:- Morning.

(Long silence)

A:- Anything that you want to discuss?

B:- No, not really.

A:- Ok.

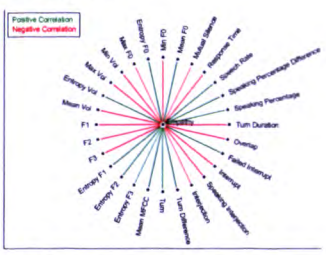
(Long silence)

B:- What about you, do you have anything to discuss?

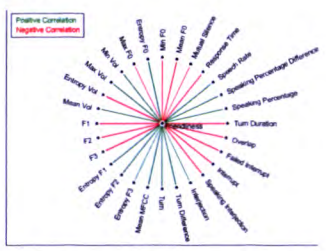
A:- No, I am just checking my emails.

.2 Correlation between social indicators and audio-visual features

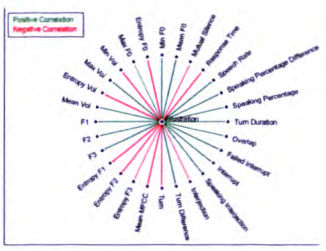
(a) Correlation between *Confusion* and audio features from AC corpus.



(b) Correlation between *Empathy* and audio features from AC corpus.



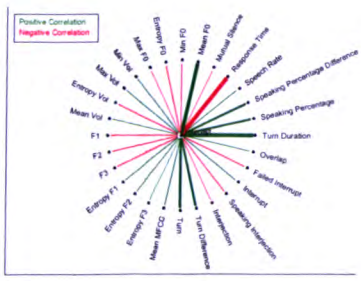
(c) Correlation between *Friendliness* and audio features from AC corpus.



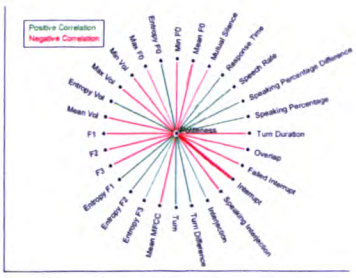
(d) Correlation between *Frustration* and audio features from AC corpus.

Figure 1: Correlation between social indicators (*Confusion*, *Empathy*, *Friendliness*, and *Frustration*) and audio features for AC corpus.

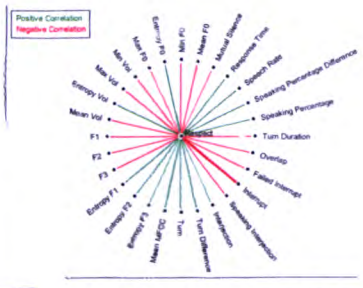
(a) Correlation between *Hostility* and audio features from AC corpus.



(b) Correlation between *Interest* and audio features from AC corpus.

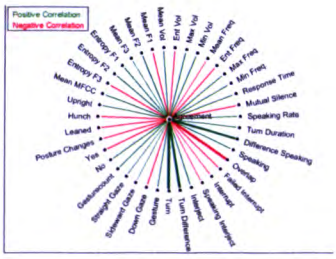


(c) Correlation between *Politeness* and audio features from AC corpus.



(d) Correlation between *Respect* and audio features from AC corpus.

Figure 2: Correlation between social indicators (*Hostility*, *Interest*, *Politeness*, and *Respect*) and audio features for AC corpus.



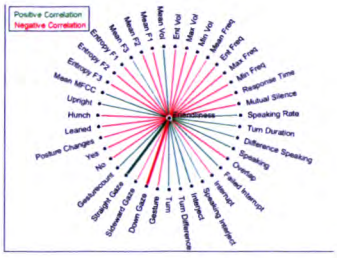
(a) Correlation between *Agreement* and audio-video features from AVC corpus.

(b) Correlation between *Confusion* and audio-video features from AVC corpus.

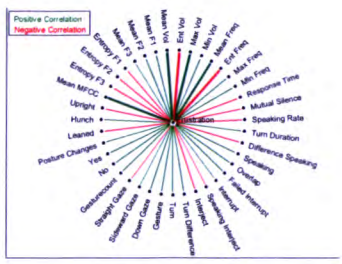
(c) Correlation between *Dominance* and audio-video features from AVC corpus.

(d) Correlation between *Empathy* and audio-video features from AVC corpus.

Figure 3: Correlation between social indicators (*Agreement*, *Confusion*, *Dominance*, and *Empathy*) and audio-video features for AVC corpus.



(a) Correlation between *Friendliness* and audio-video features from AVC corpus.



(b) Correlation between *Frustration* and audio-video features from AVC corpus.

(c) Correlation between *Politeness* and audio-video features from AVC corpus.

(d) Correlation between *Respect* and audio-video features from AVC corpus.

Figure 4: Correlation between social indicators (*Friendliness*, *Frustration*, *Politeness*, and *Respect*) and audio-video features for AVC corpus.